Published in final edited form as:
Stat Med. 2011 December 10; 30(28): 3361–3371. doi:10.1002/sim.4337.

Integrative analysis of multiple cancer prognosis studies with gene expression measurements

Shuangge Ma[a,*,†], **Jian Huang**[b], **Fengrong Wei**[c], **Yang Xie**[d], and **Kuangnan Fang**[e]

[a]School of Public Health, Yale University, New Haven, CT, USA

[b]Department of Statistics and Actuarial Science, University of Iowa, Iowa, IA, USA

[c]Department of Mathematics, University of West Georgia, Carrollton, GA, USA

[d]Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, USA

[e]Department of Statistics, Xiamen University, Xiamen, Fujian, China

Abstract

Although in cancer research microarray gene profiling studies have been successful in identifying genetic variants predisposing to the development and progression of cancer, the identified markers from analysis of single datasets often suffer low reproducibility. Among multiple possible causes, the most important one is the small sample size hence the lack of power of single studies. Integrative analysis jointly considers multiple heterogeneous studies, has a significantly larger sample size, and can improve reproducibility. In this article, we focus on cancer prognosis studies, where the response variables are progression-free, overall, or other types of survival. A group minimax concave penalty (GMCP) penalized integrative analysis approach is proposed for analyzing multiple heterogeneous cancer prognosis studies with microarray gene expression measurements. An efficient group coordinate descent algorithm is developed. The GMCP can automatically accommodate the heterogeneity across multiple datasets, and the identified markers have consistent effects across multiple studies. Simulation studies show that the GMCP provides significantly improved selection results as compared with the existing meta-analysis approaches, intensity approaches, and group Lasso penalized integrative analysis. We apply the GMCP to four microarray studies and identify genes associated with the prognosis of breast cancer.

Keywords

integrative analysis; cancer prognosis; microarray; penalized selection

1. Introduction

Cancer is a disease of the genome. The development of microarray technologies makes it possible to simultaneously profile the expressions of thousands of genes, searching for those associated with the development and progression of cancer. This article focuses on cancer prognosis studies, where the response variables are progression-free, overall, or other types of survival. Microarray studies have been conducted on the prognosis of breast cancer, ovarian cancer, lung cancer, lymphoma, and other types of cancers [1]. Significant progress

*Correspondence to: Shuangge Ma, School of Public Health, Yale University, New Haven, CT, USA.
†shuangge.ma@yale.edu

has been made. As an example, with breast cancer prognosis, several gene signatures have been validated in wet labs and are currently being tested in prospective clinical studies [2].

Despite promising successes, markers identified from the analysis of single-cancer microarray studies often suffer a lack of reproducibility. This problem has been noted in multiple studies [3–5] and can also be partly seen from our numerical study in Sections 4 and 5. There are multiple possible causes for the lack of reproducibility. First, the microarray profiling technique is still not perfect. The gene expression measurements are noisy and subject to measurement errors. Although profiling techniques have been significantly improved in the past few years, the corresponding improvement in the reproducibility of identified markers has not been observed. Second, most cancer prognosis microarray studies are retrospective and do not have strict subject selection criteria. The heterogeneity among study subjects may reduce the comparability of gene signatures. The lack of reproducibility caused by difference in study subjects can be partly improved by properly adjusting for demographic and other variables in the construction of gene signatures. Third, it is possible that multiple distinct sets of genes belong to the same pathways and/or have the same biological functions. Pathway-based and network-based studies have been conducted, which have led to improved reproducibility for some gene signatures. The last and perhaps the most important reason for the lack of reproducibility is the small sample sizes and hence lack of power of single studies—a typical cancer microarray study profiles the expressions of $\sim 10^{3-4}$ genes on $\sim 10^{2-3}$ subjects.

For the prognosis of multiple cancers including for example breast cancer, ovarian cancer, and lymphoma, there are multiple independent studies sharing comparable designs. In addition, many researchers have made raw data from their studies publicly available. Data warehouses that host cancer microarray studies include GEO, Array Express, Oncomine, and others. Thus, a cost-effective remedy for the lack of power and lack of reproducibility problem is to pool and analyze data from multiple comparable studies. A series of recent studies show that, despite the heterogeneity among studies, pooling and analyzing multiple datasets may significantly improve reproducibility [5–7].

Multi-dataset approaches can be classified as meta-analysis and integrative analysis approaches. In meta-analysis, multiple datasets are analyzed separately. Then summary statistics, for example the lists of identified markers or *p*-values, are combined across multiple datasets. Integrative analysis, in contrast, pools and analyzes raw data. A family of integrative analysis approaches, referred to as ‘intensity approaches’ in the literature, transform gene expressions from different studies and different platforms to the same reference distributions. After transformation, multiple datasets are combined and analyzed using single-dataset approaches. A significant drawback of intensity approaches is that they need to be conducted on a case-by-case basis with no generically applicable transformation. In recent studies, Ma and Huang [5], Ma et al. [8], and others developed approaches that do not require the full comparability of measurements from different studies/platforms. Those studies focus on diagnosis studies with categorical response variables under the logistic regression model. Performance of those approaches with prognosis studies has not been investigated. In addition, the approach in [5] does not have a well-defined objective function. The group bridge approach in [8] is computationally expensive.

To overcome the limitations of current genome-wide profiling studies and translate the research results into real clinical practice, we need to identify a small set of markers, use high-quality Clinical Laboratory Improvement Amendments certified assays to measure the mRNA expressions of these markers for patient samples, and then develop and validate prediction models in well-designed clinical studies. As this type of clinical studies is very expensive and time consuming, accurate identification of markers is crucial. The goal of this

study was to accurately identify a set of cancer prognosis markers using a novel integrative analysis approach which can combine information from multiple heterogeneous studies. Through simulation studies and data example, we shall show that the proposed method can significantly improve the accuracy of marker identification. As discussed in [6, 7] and others, additional complications in analysis of multiple datasets (compared with analysis of single datasets) may include data selection, interpretation of identified markers, utilization of identified markers, and others. As they are not the focus of this study, we acknowledge their importance but refer to published studies for established guidelines.

For cancer prognosis studies, we describe the relationship between survival and gene expressions using the accelerated failure time (AFT) model. Unlike the Cox or additive risk models, the AFT model describes the event time directly and may have more lucid interpretations [9]. It has been adopted in [10–13] and several other studies for modeling prognosis data with microarray measurements. For marker selection, we adopt a group minimax concave penalty (GMCP) penalization approach. Penalization approaches have been extensively adopted for variable selection with single high-dimensional datasets. The literature is too vast to be reviewed here. With single prognosis studies, available penalization approaches include Lasso, elastic net, Smoothly Clipped Absolute Deviation (SCAD), bridge, minimax concave penalty (MCP), their extensions and many others. Those approaches have been developed for the analysis of single datasets and cannot be directly applied to the analysis of multiple heterogeneous datasets. Ma et al. [8] adopts a two-norm group bridge approach for the analysis of binary classification data with logistic regression models. Its performance with survival data has not been investigated. The GMCP approach is proposed in [14] for the analysis of single datasets. This study differs significantly from [14] along the following aspects. Huang et al. [14] analyzes continuous response variables under the linear regression model, whereas this study analyzes censored survival data. In [14], the grouping structure comes from dummy variables for categorical data or clusters of covariates. In contrast, in this study, the grouping structure comes from effects of single genes in multiple studies and is naturally defined.

This study has been motivated by the prevalence of cancer prognosis studies with gene expression measurements, effectiveness of integrative analysis in improving the reproducibility of identified markers, and the significant difference in data and model settings from existing integrative analysis studies. The rest of the article is organized as follows. In Section 2, we describe the data and model settings. In Section 3, we describe marker selection using the GMCP. We also develop an effective computational algorithm. We conduct a simulation study in Section 4 and analyze four breast cancer prognosis studies in Section 5. The article concludes with discussion in Section 6.

2. Integrative analysis of multiple cancer microarray prognosis studies

The strength of meta-analysis and integrative analysis lies in their ability to borrow information across multiple datasets. To make this feasible, when pooling and analyzing multiple datasets, it is usually required that those datasets share certain common ground [7]. As the goal of this study is to identify cancer genomic markers, we focus on the scenario where multiple datasets share the same set of markers. This can be achieved by carefully evaluating and selecting datasets using for example the Minimum Information About a Microarray Experiment criteria and personal expertise. A representative example is the pancreatic cancer study in [6]. In addition, for data deposited at the National Center for Biotechnology Information, GEO datasets have been assembled by GEO staff using a collection of biologically and statistically comparable samples. With those selected studies, it is reasonable to expect that they share the same set of markers.

In microarray studies, measurements from different studies and different platforms (for example cDNA and Affymetrix) are not directly comparable, which makes directly combing multiple datasets inappropriate. There is no guarantee that cross-study (platform) normalization or transformation always exists. In addition, other confounders may alter the relationship between gene expressions and cancer outcomes. For example, for both smokers and nonsmokers, gene NET1 is a marker for the development of lung cancer. However, the strengths of associations measured with magnitudes of regression coefficients are different between the two groups.

2.1. Data and model settings

Suppose that there are M independent studies measuring the same cancer prognosis outcomes, and within each study, there are the same d gene expressions. With the pangenomic arrays becoming the routine practice, the matched gene sets can often be achieved. The discussion on partially matched gene sets is postponed to Section 4. Let $T^1, \ldots, T^M$ be the logarithms (or other known monotone transformations) of the failure times and $X^1, \ldots, X^M$ be the length d covariates (gene expressions). For $m = 1, \ldots, M$, assume the AFT model

$$T^m=\alpha^m+\beta^{m'}X^m+\varepsilon^m. \quad (1)$$

Here α^m is the unknown intercept, $\beta^m \in \mathbb{R}^d$ is the regression coefficient vector, $\beta^{m'}$ is the transpose of β^m, and ε^m is the random error with an unknown distribution. Denote $C^1, \ldots, C^M$ as the logarithms of random censoring times. Under right censoring, observations are $(Y^m, \delta^m, X^m$ for $m = 1 \ldots M$. Here $Y^m = min(T^m, C^m)$ and $\delta^m = I(T^m \le C^m)$.

To more explicitly describe the essential data and model settings, consider a hypothetical example with four independent studies and $d = 1000$ gene expressions. Assume that only the first two genes are associated with prognosis. A hypothetical set of regression coefficients are presented in Table I. The regression coefficients and corresponding statistical models have the following features. First, only the first two prognosis-associated genes have nonzero regression coefficients. That is, the models are sparse. Marker identification amounts to discriminating genes with nonzero coefficients from those with zero coefficients. Second, as the four studies share the same set of markers, the four models have the same sparsity structure. Third, to accommodate heterogeneity, the nonzero coefficients of markers are allowed to differ across studies. This strategy has been proved to be effective in [5, 15] and others.

2.2. Weighted least squares estimation

With the AFT model, popular estimation approaches include those proposed in [16,17] among others. A common drawback of those approaches is the high computational cost, which makes them unsuitable for gene expression data. A computationally more affordable approach is the weighted least squares estimation developed in [18]. Particularly, this estimation approach has been applied to gene expression data in [11,13].

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

and can be computed as $w_1^m = \frac{\delta_{(1)}^m}{n^m}$ and $w_{(i)}^m = \frac{\delta_{(i)}^m}{n^m - i + 1} \prod_{j=1}^{i-1} \left(\frac{n^m - j}{n^m - j + 1} \right)^{\delta_{(j)}^m}$ for $i = 2 \ldots n^m$. For study m, the weighted least squares objective function is defined as

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

We 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 $Y_{(i)}^{m*} = \sqrt{w_i^m} \left(Y_{(i)}^m - \frac{\sum w_i^m Y_{(i)}^m}{\sum w_i^m} \right)$. We define the overall objective function as

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

where $\beta = (\beta^1, \ldots, \beta^M)$ is the $d \times M$ regression coefficient matrix.

The objective functions R^ms are not normalized by sample size. Thus, larger datasets have more contributions. This is intuitively reasonable as larger studies have more power and thus should have more 'weights'.

3. Marker selection using group minimax concave penalty

For marker selection in analysis of single datasets, Zhang [19] proposes the MCP approach. With MCP, the penalty function is defined as

$$\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\gamma \lambda} \right)_+ dx,$$

where λ is the penalty parameter and γ is the regularization parameter. $x_+ = \max(0, x)$.

The most popular penalty is the Lasso penalty, where the penalty is a linear function of the absolute value of the regression coefficient. However, as it applies 'too much penalty' to large regression coefficients and 'too little penalty' to small regression coefficients, the Lasso tends to over-select. Under certain data and model settings, a few penalties, particularly including the bridge, SCAD, MCP, and adaptive Lasso, have been shown to have the selection consistency property. With small regression coefficients, the MCP applies the same amount of penalization as the Lasso. When regression coefficients increase, the degree of penalty decreases. When $|t| > \gamma\lambda$, the penalty drops to zero.

Consider the integrative analysis settings described in Section 2. Denote β_j^m as the jth component of β^m. $\beta_j = (\beta_j^1, \ldots, \beta_j^M)$ is the jth row of β and represents the coefficients of gene j across M studies. Define

$$\widehat{\beta}=\text{argmin}_{\beta}\left\{R(\beta)+n\sum_{j=1}^{d}\rho(\left\|\beta_j\right\|_2;\lambda,\gamma)\right\}, \quad (3)$$

where $\left\|\beta_j\right\|_2=\left((\beta_j^1)^2+\ldots+(\beta_j^M)^2\right)^{1/2}$ is the L_2 norm.

The GMCP penalty has been motivated by the following considerations. When $M = 1$ (a single dataset), the GMCP simplifies to the MCP penalty, which has been shown to have the selection consistency property [19]. In integrative analysis of multiple prognosis studies, for a specific gene, we need to evaluate its overall effects in multiple datasets. To achieve such a goal, we treat its M regression coefficients as a *group* and conduct group-level selection. When a group is selected, the corresponding gene is identified as associated with prognosis. Otherwise, it is identified as noise. Within specific groups, as genes are expected to have consistent (either all zero or all nonzero) effects across multiple studies, the L_2 group norm is adopted.

The GMCP penalization approach adopted in this study differs from the one in [13] along the following aspects. The first and most significant difference comes from the data structure. In [13], the grouping structure comes from clusters of covariates within single datasets. In contrast, in this study, one group corresponds to one gene but from multiple independent datasets. Second, Huang et al. [13] investigates continuous data, whereas we analyze censored survival data. Third, in [13], the L_2 norm is rescaled by the sample variance–covariance matrix. Such a rescaling is necessary in [13] as covariates in the same groups tend to be correlated. In contrast, in this study, different components within the same groups correspond to different *independent* datasets. Thus, in this study, we choose not to conduct the rescaling, which may make the penalized estimates more intuitive and more interpretable. In addition, unlike in [13], different groups have the same sizes—all equal to the number of independent studies. Thus, rescaling of parameter λ is not needed.

3.1. Computational algorithm

We use a group coordinate descent approach, which is a natural extension of regular coordinate descent algorithm, to compute the proposed GMCP estimate. In analysis of single datasets, the coordinate descent algorithm has been extensively used for computing penalized estimates [20]. The group coordinate descent algorithm is the integrative analysis counterpart of the algorithm described in [21] and proceeds as follows.

Algorithm

1. Initialize $\hat{\beta} = 0$;
2. for $j = 1, \ldots, d$,
    a. With the current estimate $\hat{\beta}$, define $\hat{\beta}(j)$, which is a $d \times M$ matrix with its kth row $\hat{\beta}(j)_k = \hat{\beta}_k$ for $k \neq j$. The jth row of $\hat{\beta}(j)_j = b = (b_1, \ldots, b_M)$ is the vector of unknown regression coefficients of gene j.
    b. Compute $\hat{\beta} = \text{argmin}\{R(\hat{\beta}(j)) + n\rho(\|b\|_2; \lambda, \gamma)\}$;
    c. Update $\hat{\beta}_j = \hat{b}$;
3. Repeat step 2 until convergence. In numerical study, we use the L_2 norm of the difference between two consecutive $\hat{\beta} < 0.01$ as the stopping rule. With our simulated and breast cancer data, convergence is achieved within 20 iterations.

The above algorithm only involves iterative computations of the marginal GMCP estimates, which can be obtained as follows. Denote $\tilde{b} = \text{argmin} R(\hat{\beta}(j))$. Note that because usually $n \gg M$ and because of the simple least squared format of R, $\tilde{b}$ can be easily obtained. The marginal GMCP estimate is then

$$\widehat{b}=\begin{cases} 0 & \text{if } \left\|\tilde{b}\right\|_2 \le \lambda \\ \frac{\gamma}{\gamma-1}\left(1-\frac{\lambda}{\left\|\tilde{b}\right\|_2}\right)\tilde{b} & \text{if } \lambda \le \left\|\tilde{b}\right\|_2 \le \lambda\gamma \\ \tilde{b} & \text{if } \left\|\tilde{b}\right\|_2 \ge \lambda\gamma. \end{cases} \quad (4)$$

The above procedure only involves simple calculations. Thus, the proposed algorithm, although iterative, is computationally affordable.

3.2. Tuning parameter selection

The GMCP involves two tuning parameters, λ and γ, which jointly determine properties of the GMCP estimates. Specifically, with a fixed γ, a larger value of λ leads to fewer genes identified as associated with prognosis. With a fixed λ, as $\gamma \to \infty$, the proposed GMCP estimates converge to group Lasso-type estimates, as can be seen from the definition of the penalty. As $\gamma \to 0$, the GMCP estimates converge to AIC/BIC-type estimates. In our numerical study, we adopt V-fold cross validation for tuning parameter selection. For λ, we search over the discrete grid of $2^{\cdots, -1, -0.5, 0, 0.5, 1, \cdots}$. For γ, we search over the discrete grid of 1.5, 2.0, …, 5.5, 6. We have numerically experimented with alternative tuning parameter selection approaches, including BIC, AIC, and leave-one-out cross validation (results omitted). We find that other tuning parameter selection techniques do not significantly outperform V-fold cross validation. We choose V-fold cross validation because of its computational simplicity.

4. Simulation study

For simplicity of notation, we have assumed matched gene sets across multiple studies. When different sets of genes are measured in different studies, we use the following rescaling approach. Assume that gene 1 is measured only in the first $K(< M)$ studies. We set $\beta_1^{K+1} = \ldots = \beta_1^M = 0$ and replace $\rho(\|\beta_1\|_2; \lambda, \gamma)$ with $\rho(\left\|\beta_1\right\|_2; \lambda, \gamma) \times \sqrt{M/K}$. The proposed approach and computational algorithm are then applicable with minor modifications.

Our simulation settings closely mimic pharmacogenomic studies, where genes have the pathway structures. Genes within the same pathways tend to have correlated expressions, whereas genes within different pathways tend to have weakly correlated or independent expressions. Among large number pathways, only a few are associated with the responses. Within those important pathways, there are some important genes, and the rest are noises. Specifically, we simulate data for four independent studies, each with 100 subjects. We simulate 50 or 100 gene clusters, with 20 genes in each cluster. Thus, the total number of gene expressions simulated is 1000 or 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 $\max(0, 1 - |j - k| \times \rho)$; 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 ρ values (weak and strong correlations). Within each of the first four clusters, there are five genes associated with the responses. There are thus a total of 20 important genes, and the rest are

noises. For important genes, we generate their regression coefficients from Unif[0.5, 1]. Ten per cent of important and noisy genes are only measured in two studies. We note that for a specific important gene, its regression coefficients in the four studies are separately simulated. Thus, with probability 1, they are not equal, reflecting the heterogeneity across studies. We generate the log event time from the AFT model with intercept equal to 0. The censoring time is generated independent of event. We adjust the censoring time distribution so that the censoring rate is ~50%.

To better gauge performance of the proposed approach, we also consider the following alternatives. (i) Meta-analysis. We first analyze each dataset separately. Genes that are identified in at least one study are identified in meta-analysis. An alternative is to consider genes identified in all four studies. However, we have examined all simulation settings and found that there are only a few such genes. When analyzing each dataset, we consider both the Lasso and MCP; (ii) an intensity approach. Since all four datasets are generated under similar settings, we adopt an intensity approach, make transformations of gene expressions, combine the four datasets, and analyze as if they were from a single study [22]. For the combined dataset, we analyze using both the Lasso and MCP; and (iii) integrative analysis with group Lasso (GLasso). The proposed GMCP approach is the integrative analysis counterpart of the MCP. Following a similar strategy, it is possible to extend the Lasso penalty to its integrative analysis counterpart—the GLasso penalty. With all the six approaches, we select the tuning parameters using fourfold cross validation. Different approaches have different ways of accommodating the heterogeneity across studies. With meta-analysis, only the lists of identified genes are pooled, which are expected to be consistent across studies. Intensity approaches attempt to remove the heterogeneity via transformations prior to analysis. In contrast, integrative analysis approaches accommodate the heterogeneity by allowing for different regression coefficients across studies.

Simulation suggests that the GMCP approach is computationally affordable. Its computational cost is comparable with that of meta-analysis and lower than that of intensity approaches. With GMCP, analysis of one replicate (with 2000 genes) takes about 3 min on a desktop PC. Summary statistics on the numbers of genes identified and true positives based on 200 replicates are shown in Table II. We can see that, although the meta-analysis approaches can identify the majority or all of the true positives, they also identify a large number of false positives. The intensity approaches can significantly outperform the meta-analysis approaches. The satisfactory performance of intensity approaches is not surprising, considering that the four simulated datasets are very similar to each other—the degree of similarity is higher than that encountered in practical data analysis. The integrative analysis approaches outperform alternatives by identifying the majority or all of true positives and a smaller number of false positives. Among the integrative analysis approaches, the GMCP significantly outperforms the GLasso by identifying a much smaller number of false positives, at the price of a very small number of false negatives. We have also experimented with a few other simulation settings and reached similar conclusions.

5. Analysis of breast cancer prognosis studies

Among women in the USA, breast cancer is the most commonly diagnosed malignancy after skin cancer and is the second leading cause of cancer deaths after lung cancer. According to the American Cancer Society, in 2009, an estimated 192,370 new cases of breast cancer were diagnosed, and 40,160 died from breast cancer. Women in the USA have a one in eight lifetime risk of developing invasive breast cancer and a one in 33 overall chance of dying from it. Biomedical studies suggest that genomic measurements may have independent predictive power for breast cancer prognosis [1, 2]. Multiple gene profiling studies have been conducted, searching for genomic measurements with predictive power for breast

cancer prognosis. We collect and analyze four breast cancer prognosis studies with microarray measurements. The same datasets have been analyzed in [4, 23]. Analysis in this study differs significantly from that in [4,23]. Specifically, the previous studies investigate the marginal effects by analyzing genes or pathways separately. In contrast, in this study, we assume that prognosis is associated with the *combined effects of multiple genes*.

We provide brief descriptions of the four studies in Table III and refer to the original publications for more detailed information. Among the four datasets, two used cDNA, one used oligonucleotide arrays, and one used Affymetrix genechips for profiling. We first conduct normalization of gene expressions for each dataset separately. With Affymetrix chips, the measurements are log2 transformed. We impute missing expressions with means across samples. We then standardize each gene expression to have zero mean and unit variance. The proposed approach does not require the direct comparability of measurements from different studies. Additional transformations, which are necessary for intensity approaches, are not needed. We match genes in the four studies using their Unigene Cluster IDs. Although the proposed approach can accommodate partially matched gene sets, for reliability, we focus on the 2555 genes that are measured in all four studies.

With the GMCP approach, 13 genes are identified as associated with breast cancer prognosis (Table IV). Searching literature suggests that several of the identified genes may have sound biological implications, which may partly support the effectiveness of the proposed approach. Particularly, gene RLF has been shown to be in fusion with gene MYCL1, which is an established marker for multiple cancers particularly including breast cancer [24]. Gene IRAK1 encodes the interleukin-1 receptor-associated kinase 1, one of the two putative serine/threonine kinases that become associated with the interleukin-1 receptor (IL1R) upon stimulation. Its involvement in breast cancer development has been investigated in [25]. The protein encoded by gene RNF14 contains a RING zinc finger, a motif known to be involved in protein–protein interactions. This protein interacts with androgen receptor (AR) and functions as a coactivator that induces AR target gene expression. A dominant negative mutant of this gene has been demonstrated to inhibit the AR-mediated growth of cancer. The protein encoded by gene GLS is the major enzyme yielding glutamate from glutamine. Significance of this enzyme derives from its implication in behavior disturbances in which glutamate acts as a neurotransmitter. Its implication in breast cancer has been discussed in [26] and references therein. Adenine phosphoribosyltransferase belongs to the purine/pyrimidine phosphoribosyltransferase family. A conserved feature of this gene is the distribution of CpG dinucleotides. This enzyme catalyzes the formation of adenosine monophosphate and inorganic pyrophosphate from adenine and 5-phosphoribosyl-1-pyrophosphate. It also produces adenine as a by-product of the polyamine biosynthesis pathway. Its implication in breast cancer is inferred in [27]. In [28], gene GSN is found co-expressed with ALCAM (activated leukocyte cell adhesion molecule), which is overexpressed in many mammary tumors. The protein encoded by gene RAD50 is highly similar to *Saccharomyces cerevisiae* Rad50, a protein involved in DNA double-strand break repair. This protein, cooperating with its partners, is important for DNA double-strand break repair, cell cycle checkpoint activation, telomere maintenance, and meiotic recombination. Knockout studies of the mouse homolog suggest that this gene is essential for cell growth and viability. Gene PIGC encodes an endoplasmic reticulum associated protein that is involved in glycosylphosphatidylinositol (GPI) lipid anchor biosynthesis. The GPI lipid anchor is a glycolipid found on many blood cells and serves to anchor proteins to the cell surface. Chitotriosidase, encoded by gene CHIT1, is secreted by activated human macrophages and is markedly elevated in the plasma of Gaucher disease patients. It is expressed and secreted by several types of solid tumors including glioblastoma, colon cancer, breast cancer and malignant melanoma. Vasodilator-stimulated phosphoprotein (VASP) is a member of the Ena-VASP protein family. 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.

We note that the estimated regression coefficients are in general small. This is mainly caused by the 'small event times' after the logarithm transformation. In addition, as other penalization approaches, the GMCP has the shrinkage property. It is possible to re-estimate the regression coefficients using only the identified markers to partly release the shrinkage.

Besides the GMCP, we also analyze data using the alternative approaches described in Section 4 and present summary results in Table V. More details are available from the authors. As seen in simulation, meta-analysis approaches identify a relatively large number of genes, with small overlap among the sets of genes identified in different datasets. Both the intensity and integrative analysis approaches identify a small number of genes. The genes identified using the proposed approach can differ significantly from those identified using alternatives. The number of overlapped genes ranges from 1 to 12.

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. It is expected that if the identified markers are more meaningful, prediction using those markers is more accurate. Specifically, we first split each dataset randomly into a training set and a testing set, with sizes 3:1. We carry out the GMCP estimation (which involves tuning parameter selection via cross validation and penalized estimation and marker selection) with the training set only and then make prediction for subjects in the testing set. Based on the predicted $\hat{\beta}^{m'} X^m$, we generate two risk groups with equal sizes. The logrank statistic is computed to evaluate the difference between survival of the two groups. For each random split, we compute the mean logrank statistics over four datasets. To avoid an extreme split, we repeat the whole process 50 times, compute the mean logrank statistics, and present the results in Table V. The proposed approach has the best prediction performance, with the logrank statistic equal to 6.576.

6. Discussion

In cancer prognosis studies with gene expression measurements, markers identified from analysis of single datasets often suffer low reproducibility because of small sample sizes and hence lack of power. Several published studies suggest that pooling and analyzing multiple studies with comparable designs may improve power and reproducibility. In this study, with multiple heterogeneous cancer prognosis studies, we adopt a GMCP penalization approach for marker selection. The proposed approach is intuitive and has affordable computational cost. Numerical studies, including simulation and analysis of breast cancer prognosis studies, suggest satisfactory performance of the proposed approach.

When modeling survival, we adopt the AFT model, which is one of the most extensively used survival models. It is of interest to investigate similar penalized integrative analysis and marker selection with alternative models such as the Cox model. However, single-dataset studies suggest that such an extension is highly nontrivial and warrants separate investigation. The proposed penalty has been motivated by the MCP penalty for the analysis of single datasets. When analyzing single datasets, a few other penalties, including the bridge, SCAD, and adaptive Lasso, have the selection consistency properties. We conjecture that it is possible to develop the integrative analysis counterparts of those penalties. As analysis of single datasets does not suggest the superiority of any penalty over the MCP, we will not further discuss other penalties. In this study, we investigate performance of the proposed approach via extensive numerical studies. With a single dataset, continuous response variable (no censoring), and simple linear regression model, Huang et al. [14] show

that asymptotic properties of the GMCP penalty can be extremely difficult to establish. The present data setting is significantly more complicated than that in [14] because of censoring. We postpone asymptotic studies to future research. Simulation study clearly establishes the superiority of the proposed approach. We choose MCP for comparison in meta-analysis and intensity approach because its penalization framework is closest to the proposed one. Lasso is also compared as it has been used as a benchmark in many studies. We acknowledge that a large number of alternative penalties can also be used in meta-analysis and intensity approach. However, as their performance is expected to be comparable with or inferior to that of MCP, we focus on MCP and Lasso for comparison. In analysis of breast cancer data, the proposed GMCP identifies the shortest list of genes, which can lead to more focused hypothesis for future validation studies and hence may be preferred. The satisfactory prediction performance partly supports the validity of proposed model and marker selection approach. With high dimensional prognosis data, there is still no well-established model diagnostics tool. Thus, model checking is not conducted. The identified markers need to be validated in independent studies before any clinical usage.

Acknowledgments

We would like to thank the editor, associate editor, and reviewer for careful review and insightful comments, which have led to a significant improvement of the paper. This study has been supported by awards LM009828, CA120988, CA152301 and CA142774 from NIH.

References

1. Knudsen, S. Cancer Diagnostics with DNA microarrays. John Wiley and Sons; 2006.
2. Cheang M, van de Rijn M, Nielsen TO. Gene expression profiling of breast cancer. Annual Review of Pathology: Mechanisms of Disease. 2008; 3:67–97.
3. Choi H, Shen R, Chinnaiyan AM, Ghosh D. A latent variable approach for meta analysis of gene expression data from multiple microarray experiments. BMC Bioinformatics. 2007; 8:364. [PubMed: 17900369]
4. Shen R, Ghosh D, Chinnaiyan AM. Prognostic meta signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics. 2004; 5:94. [PubMed: 15598354]
5. Ma S, Huang J. Regularized gene selection in cancer microarray meta-analysis. BMC Bioinformatics. 2009; 10:1. [PubMed: 19118496]
6. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schachert H, Kloppel G, Saeger H, Pilarsky C. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. Oncogene. 2005; 24:5079–5088. [PubMed: 15897887]
7. Guerra, R.; Goldstein, DR. Chapman and Hall/CRC; 2009.
8. Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional datasets. Biostatistics. In press.
9. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in Medicine. 1992; 11:1871–1879. [PubMed: 1480879]
10. 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; 63:259–271. [PubMed: 17447952]
11. Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high dimensional covariates. Biometrics. 2006; 62:813–820. [PubMed: 16984324]
12. Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. BMC Bioinformatics. 2009; 9:269. [PubMed: 18538026]
13. Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge penalty. Lifetime Data Analysis. 2010; 16:176–195. [PubMed: 20013308]
14. Huang J, Wei F, Ma S. Consistent group selection and estimation via normed minimax concave penalty. 2010 Unpublished manuscript.

15. Stevens JR, Doerge RW. Meta-analysis combines Affymetrix microarray results across laboratories. Comparative and Functional Genomics. 2005; 6:116–122. [PubMed: 18629222]
16. Buckley J, James I. Linear regression with censored data. Biometrika. 1979; 66:429–436.
17. Ying ZL. A large sample study of rank estimation for censored regression data. Annals of Statistics. 1993; 21:76–99.
18. Stute W. Consistent estimation under random censorhip when covariables are available. Journal of Multivariate Analysis. 1993; 45:89–103.
19. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics. 2010; 38:894–942.
20. Friedman J, Hastie T, Tibshirani R. Regularized paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33(1):1–22. [PubMed: 20808728]
21. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics. 2010; 5:232–253. [PubMed: 22081779]
22. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene expression studies via cross platform normalization. Bioinformatics. 2008; 24:1154–1160. [PubMed: 18325927]
23. Ma S, Kosorok MR. Detection of gene pathways with predictive power for breast cancer prognosis. BMC Bioinformatics. 2010; 11:1. [PubMed: 20043860]
24. Bhatti P, Doody MM, Rajaraman P, Alexander BH, Yeager M, Hutchinson A, Burdette L, Thomas G, Hunter DJ, Simon SL, Weinstock RM, Rosenstein M, Stovall M, Preston DL, Linet MS, Hoover RN, Chanock SJ, Sigurdson AJ. Novel breast cancer risk alleles and interaction with ionizing radiation among U.S. radiologic technologists. Radiation Research. 2010; 173:214–224. [PubMed: 20095854]
25. Pilarsky C, Wenzig M, Specht T, Saeger HD, Grutzmann R. Identification and validation of commonly overexpressed genes in solid tumors by comparison of microarray data. Neoplasia. 2004; 6:744–750. [PubMed: 15720800]
26. Kaufmann Y, Todorova VK, Luo S, Klimberg VS. Glutamine affects glutathione recycling enzymes in a DMBA-induced breast cancer model. Nutrition and Cancer. 2008; 60:518–525. [PubMed: 18584486]
27. Lubin M, Lubin A. Selective killing of tumors deficient in methylthioadenosine phosphorylase: a novel strategy. PLoS One. 2009; 29:e5735. [PubMed: 19478948]
28. Hein S, Muller V, Kohler N, Wikman H, Krenkel S, Streichert T, Schweizer M, Riethdorf S, Assmann V, Ihnen M, Beck K, Issa R, Janicke F, Pantel K, Milde-Langosch K. Biologic role of activated leukocyte cell adhesion molecule overexpression in breast cancer cell lines and clinical tumor tissue. Breast Cancer Research and Treatment. 2010 In press.
29. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein LP, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Science. 2001; 98:10869–10874.
30. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, 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; 415:530–536. [PubMed: 11823860]
31. 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; 361:1590–1596. [PubMed: 12747878]
32. 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. Proceedings of the National Academy of Science. 2003; 100:10393–10398.

Table I

Matrix of regression coefficients for a hypothetical study with four datasets and 1000 genes. Only the first two genes are associated with prognosis.

	Dataset			
Gene	**D1**	**D2**	**D3**	**D4**
1	0.20	0.05	0.13	0.27
2	−0.11	−0.17	−0.12	−0.21
3	0	0	0	0
…		…		
999	0	0	0	0
1000	0	0	0	0

Table II

Simulation results based on 200 replicates. Correlation structures include auto-regressive (Auto), banded (Band), and compound symmetry (Comp). P, number of covariates identified; TP, number of true positives.

				Meta analysis				Intensity approach				Integrative analysis			
				Lasso		MCP		Lasso		MCP		GLasso		GMCP	
Sample	#Cov	Corr	ρ	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP
100	1000	Auto	0.3	171	19	95	20	136	20	20	20	104	20	20	20
			0.7	144	20	80	19	125	20	24	20	41	20	20	17
		Band	0.2	129	20	64	19	122	20	36	20	36	20	24	18
			0.33	152	20	72	20	128	20	21	20	60	20	29	17
		Comp	0.3	154	18	101	20	112	20	20	20	150	20	22	20
			0.7	109	20	89	18	105	20	41	19	64	20	26	19
100	2000	Auto	0.3	244	19	153	20	176	20	22	20	178	20	20	20
			0.7	253	19	128	20	188	20	22	20	177	20	20	20
		Band	0.2	252	19	141	20	174	20	36	20	180	20	20	20
			0.33	254	19	158	20	164	20	21	20	181	20	20	20
		Comp	0.3	242	20	142	20	98	20	22	20	182	20	20	20
			0.7	257	20	140	20	189	20	53	20	174	20	20	20

Table III

Breast cancer prognosis studies.

Reference	Platform	Gene	Sample
Sorlie et al. [29]	cDNA	8102	58
van't Veer et al. [30]	Oligonucleotide	24,481	78
Huang et al. [31]	Affymetrix	12,625	71
Sotiriou et al. [32]	cDNA	7650	98

Table IV

Analysis of breast cancer prognosis studies. Genes identified using the GMCP and their estimates.

Unigene	Gene Name	D1	D2	D3	D4
Hs.13321	Rearranged L-myc fusion (RLF)	−0.015	−0.005	−0.004	−0.003
Hs.182018	Interleukin-1 receptor-associated kinase 1 (IRAK1)	−0.003	−0.004	−0.009	−0.001
Hs.215857	Ring finger protein 14 (RNF14)	0.003	0.001	0.003	0.001
Hs.239189	Glutaminase (GLS)	−0.004	0.002	0.008	−0.004
Hs.25363	Presenilin 2 (Alzheimer disease 4) (PSEN2)	0.018	0.001	0.01	0.001
Hs.28914	Adenine phosphoribosyltransferase (APRT)	−0.014	−0.004	−0.004	−0.003
Hs.290070	Gelsolin (GSN)	−0.015	−0.005	−0.002	−0.002
Hs.41587	RAD50 homolog (*S. cerevisiae*) (RAD50)	0.001	0.001	0.001	0.001
Hs.433030	Phosphatidylinositol glycan anchor biosynthesis, class C (PIGC)	−0.004	−0.001	−0.004	−0.002
Hs.75584	Exosome component 10 (EXOSC10)	−0.001	−0.001	−0.001	−0.001
Hs.80768	Chloride channel 7 (CLCN7)	0.004	0.001	0.008	0.001
Hs.91093	Chitinase 1 (chitotriosidase) (CHIT1)	0.003	−0.004	0.018	−0.003
Hs.93183	Vasodilator-stimulated phosphoprotein (VASP)	−0.014	−0.003	−0.007	−0.002

Table V

Analysis of breast cancer prognosis studies. With meta analysis approaches, numbers in the '()' are the number of genes identified with each individual datasets. Overlap: number of overlapped genes with those identified with GMCP.

Approach		Gene	Overlap	Logrank
Meta-analysis	Lasso	81 (25, 20, 24, 13)	10	2.661
	MCP	59 (10, 13, 16, 21)	8	1.612
Intensity approach	Lasso	32	2	1.884
	MCP	24	1	3.849
Integrative analysis	GLasso	42	12	2.100
	GMCP	13	–	6.576