Published in final edited form as:

Biometrics. 2014 September ; 70(3): 480–488. doi:10.1111/biom.12177.

Integrative Analysis of Prognosis Data on Multiple Cancer Subtypes

Jin Liu¹, Jian Huang², Yawei Zhang³, Qing Lan⁴, Nathaniel Rothman⁴, Tongzhang Zheng³, and Shuangge Ma^{3,*}

¹Division of Epidemiology and Biostatistics, UIC School of Public Health

²Department of Statistics & Actuarial Science, University of Iowa

³School of Public Health, Yale University

⁴Division of Cancer Epidemiology and Genetics, NCI, NIH

Summary

In cancer research, profiling studies have been extensively conducted, searching for genes/SNPs associated with prognosis. Cancer is diverse. Examining the similarity and difference in the genetic basis of multiple subtypes of the same cancer can lead to a better understanding of their connections and distinctions. Classic meta-analysis methods analyze each subtype separately and then compare analysis results across subtypes. Integrative analysis methods, in contrast, analyze the raw data on multiple subtypes simultaneously and can outperform meta-analysis methods. In this study, prognosis data on multiple subtypes of the same cancer are analyzed. An AFT (accelerated failure time) model is adopted to describe survival. The genetic basis of multiple subtypes is described using the heterogeneity model, which allows a gene/SNP to be associated with prognosis of some subtypes but not others. A compound penalization method is developed to identify genes that contain important SNPs associated with prognosis. The proposed method has an intuitive formulation and is realized using an iterative algorithm. Asymptotic properties are rigorously established. Simulation shows that the proposed method has satisfactory performance and outperforms a penalization-based meta-analysis method and a regularized thresholding method. An NHL (non-Hodgkin lymphoma) prognosis study with SNP measurements is analyzed. Genes associated with the three major subtypes, namely DLBCL, FL, and CLL/SLL, are identified. The proposed method identifies genes that are different from alternatives and have important implications and satisfactory prediction performance.

Keywords

Cancer prognosis; Integrative analysis; Genetic association; Marker identification; Penalization

*shuangge.ma@yale.edu.

Supplementary Materials

Web Appendices, Tables, and Figure referenced in Sections 3–5 and R code are available with this paper at the *Biometrics* website on Wiley Online Library.

1. Introduction

Profiling studies have been extensively conducted in cancer research, searching for markers such as SNPs and genes that are associated with prognosis. Cancer is diverse. Different subtypes of the same cancer usually have different prognosis patterns and different associated genes/SNPs. Development in this article has been partly motivated by the analysis of data on multiple subtypes of NHL (non-Hodgkin lymphoma), which is a heterogeneous group of malignancies ranging from very indolent forms to aggressive ones. Different subtypes of NHL are largely different (Zhang et al. 2011). On the other hand, there are evidences that they share common susceptibility genes. More details are provided in Section 5. Another setting where the proposed method is also applicable is the analysis of multiple types of cancers. Susceptibility genes shared by multiple subtypes of the same cancer or multiple types of cancers represent the more essential features of cancer, whereas type-specific genes determine the uniqueness of different cancers (Rhodes et al. 2004; Goh and Choi 2012).

With data on multiple subtypes of the same cancer, most existing methods analyze each subtype separately and then compare results across subtypes. For NHL, see for example Han et al. (2010) and Ma et al. (2010). Such a strategy fits the classic meta-analysis framework. With high-dimensional measurements such as SNPs, data on individual subtypes have the “large d , small n ” characteristic, with the sample size n much smaller than the number of SNPs d . Because of low sample size, susceptibility genes/SNPs identified from the analysis of individual subtypes may have unsatisfactory properties. Recent studies have shown that, when multiple datasets (multiple subtypes in this study) have overlapped susceptibility SNPs/genes, integrative analysis methods simultaneously analyze the raw data of multiple datasets and outperform single-dataset and classic meta-analysis methods (Liu et al. 2012; Ma et al. 2009; Ma et al. 2012).

Our goal is to analyze data on multiple cancer subtypes and identify genes associated with prognosis. The genetic basis of multiple datasets can be described using two models. The homogeneity model assumes that if a functional unit (gene or SNP) is identified, it is identified as associated with prognosis in all datasets (Liu et al. 2012). The heterogeneity model, on the other hand, allows a gene or SNP to be associated with prognosis in some datasets but not others and is more appropriate for the present data settings. Because of its complexity, the heterogeneity model has been less investigated. Compared with the existing studies especially those on gene expression data, the present one has additional complexity. Multiple SNPs may correspond to the same gene, and it is important to accommodate the “SNP-with-gene” structure and allow SNPs corresponding to the same gene to have different effects. The only available method that is tailored to the data settings analyzed in this study is Ma et al. (2012), which adopts thresholding for marker identification. The thresholding method does not have a well-defined objective function, and its properties have not been established. In addition, it has more tuning parameters than the penalization approach adopted in this study. Another advancement of this study is the analysis of NHL prognosis data, which may provide insights into the genetic basis of this deadly disease.

Integrative analysis of data on multiple cancer subtypes is challenging. In some studies, the subtype information may be partial or even wrong. In addition, the definition of subtypes is still evolving. For NHL, we refer to Zhang et al. (2011) for relevant discussions. When there are a lot of subtypes, the set of subtypes chosen for analysis needs to be determined by the scientific question of interest, quality of data, sample size, evidence from epidemiologic studies, and other factors. We acknowledge the importance and difficulty of these issues. In this study, we focus on the development of a new analysis method and refer to other publications for relevant discussions.

2. Integrative Analysis under the Heterogeneity Model

2.1 AFT model and weighted least squares estimation

Assume that there are M subtypes of the same cancer, and there are n^m iid observations for subtype m . The total sample size is $n = \sum_{m=1}^M n^m$. For subtype m , denote T^m as the logarithm of failure time. Denote $\mathbf{X}_o^m$ as the length- d covariate vector. The subscript “ o ” is used to discriminate the “original” against “weighted” covariates (to be defined below). For now, assume that the same set of covariates is measured for all subtypes. In Section 3, rescaling is used to accommodate partially matched covariate sets.

For subject i of subtype m , assume an AFT (accelerated failure time) model

$T_i^m = \beta_0^m + \mathbf{X}_{oi}^m \boldsymbol{\beta}^m + \varepsilon_i^m$, where β_0^m is the intercept, $\boldsymbol{\beta}^m \subseteq \mathbb{R}^d$ is the vector of regression coefficients, and ε_i^m is the error term. Under right censoring, we observe $(Y_{oi}^m, \delta_i^m, \mathbf{X}_{oi}^m)$, where $Y_{oi}^m = \min\{T_i^m, C_i^m\}$, C_i^m is the logarithm of censoring time, and $\delta_i^m = I\{T_i^m \leq C_i^m\}$ is the event indicator.

Let $\hat{F}^m$ be the Kaplan-Meier estimator of the distribution function F^m of T^m .

$\hat{F}^m(y) = \sum_{i=1}^{n^m} \omega_i^m 1\{Y_{oi}^m \leq y\}$, where ω_i^m s are the jumps in the Kaplan-Meier estimator

and can be computed as $\omega_1^m = \frac{\delta_{(1)}^m}{n^m}$ and $\omega_i^m = \frac{\delta_{(i)}^m}{n^m - i + 1} \prod_{j=1}^{i-1} \left(\frac{n^m - j}{n^m - j + 1} \right)^{\delta_{(j)}^m}$ for $i = 2, \dots, n^m$.

$Y_{o(1)}^m \leq \dots \leq Y_{o(n^m)}^m$ are the order statistics of Y_{oi}^m s, and $\delta_{(1)}^m, \dots, \delta_{(n^m)}^m$ are the associated event indicators. Similarly, let $\mathbf{X}_{o(1)}^m, \dots, \mathbf{X}_{o(n^m)}^m$ be the associated covariate vectors of the ordered Y_{oi}^m s. Stute (1996) proposed minimizing the weighted least squares objective

function $L^m(\beta_0^m, \boldsymbol{\beta}^m) = \frac{1}{2} \sum_{i=1}^{n^m} \omega_i^m (Y_{oi}^m - \beta_0^m - \mathbf{X}_{oi}^m \boldsymbol{\beta}^m)^2$.

Let

$\bar{\mathbf{X}}_\omega^m = \sum_{i=1}^{n^m} \omega_i^m \mathbf{X}_{oi}^m / \sum_{i=1}^{n^m} \omega_i^m$, $\bar{Y}_\omega^m = \sum_{i=1}^{n^m} \omega_i^m Y_{oi}^m / \sum_{i=1}^{n^m} \omega_i^m$, $\mathbf{X}_{\omega(i)}^m = (\omega_i^m)^{1/2} (\mathbf{X}_{oi}^m - \bar{\mathbf{X}}_\omega^m)$

, and $Y_{\omega(i)}^m = (\omega_i^m)^{1/2} (Y_{oi}^m - \bar{Y}_\omega^m)$. Using the weighted centered values, the intercept is zero.

The objective function is then $L^m(\boldsymbol{\beta}^m) = \frac{1}{2} \sum_{i=1}^{n^m} (Y_{\omega(i)}^m - \mathbf{X}_{\omega(i)}^m \boldsymbol{\beta}^m)^2$. Assume independence

between data for the M subtypes. The overall objective function is $L(\boldsymbol{\beta}) = \sum_{m=1}^M L^m(\boldsymbol{\beta}^m)$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^M)'$.

The AFT family contains a large number of models with different error distributions. Here, to be flexible, we assume unknown error distributions. Among the available estimation approaches, the Stute's approach has a simple weighted least squares form and the lowest computational cost. Low-dimensional studies suggest that different estimation approaches have different advantages, with no one dominating the others. For high-dimensional data, there is a lack of model diagnostics tools. The AFT model is adopted because of its simple form. The Stute's estimation is adopted because of computational simplicity. We defer the comparison of different models and estimation approaches to future studies.

2.2 Heterogeneity model

With the significantly different prognosis patterns of different subtypes, the homogeneity model can be too restricted. The heterogeneity model allows the sets of susceptibility genes/SNPs to be different across subtypes, includes the homogeneity model as a special case, and can be more flexible.

Consider a study with three subtypes and eight SNPs corresponding to four genes (Table 1). Gene 1 is associated with the prognosis of all three subtypes; Gene 2 is associated with the first two subtypes but not the third one; Gene 3 is associated with only the third subtype; And gene 4 is not associated with any subtype. In Table 1, we show the regression coefficients whose main characteristics reflect the essence of integrative analysis under the heterogeneity model. Unimportant genes/SNPs not associated with prognosis have no effect and zero regression coefficient. Penalized marker selection amounts to identifying the sparsity structure of models, that is, discriminating nonzero regression coefficients from zero ones. For an important gene/SNP (for example SNP 1_1), its strengths of association with multiple subtypes, which are measured with regression coefficients, can be different for different subtypes. In this study, our goal is to identify important genes that contain prognosis-associated SNPs. Within an important gene, no further selection is conducted. Thus, SNPs within the same gene have the "all in or all out" property. Such a strategy has been advocated by Ma et al. (2012). An alternative strategy is to identify important genes as well as important SNPs within the selected genes. However, as to be discussed in Section 3.2, it may bring much complexity and high computational cost.

3. Marker Identification

3.1 Penalized estimation

Assume that the d SNPs belong to J genes. To accommodate partially matched gene sets, without loss of generality, assume that gene j is measured only for the first M_j subtypes. Denote d_{jm} as the number of SNPs in the j th gene and m th subtype with coefficient vector $\beta_j^m = (\beta_{j1}^m, \dots, \beta_{jd_{jm}}^m)'$. The subscript m is kept to accommodate partially matched SNP sets for the same gene. Then $\beta_j = (\beta_j^1, \dots, \beta_j^{M_j})'$ contains the regression coefficients for all SNPs in gene j across all subtypes. Here notations are more complicated than those in Section 2 to accommodate the "SNP-within-gene" structure and partially matched SNP/gene sets. Denote $\beta = (\beta_1', \dots, \beta_J')'$.

Consider the penalized estimate

$$\hat{\beta} = \operatorname{argmin} \{L(\beta) + P_{\lambda_n, \gamma}(\beta)\}.$$

A nonzero component of $\hat{\beta}$ indicates an association between the corresponding gene (SNP) and subtype's prognosis. Consider the penalty function

$$P_{\lambda_n, \gamma}(\beta) = \lambda_n \sum_{j=1}^J c_j \left(\sum_{m=1}^{M_j} \sqrt{d_{jm}} \|\beta_j^m\| \right)^\gamma, \quad (1)$$

where $\lambda_n > 0$ is a data-dependent tuning parameter, $c_j \propto M_j^{1-\gamma}$ is a constant and accommodates partially matched gene sets, $\|\cdot\|$ is the L_2 norm, and $0 < \gamma < 1$ is the fixed bridge parameter. Statistical properties of the estimate are established in Supplementary Materials.

The above penalty has been designed to tailor our special data and model characteristics. In our analysis, genes are the basic functional units. The penalty is the sum of J individual terms, with one for each gene. For a specific gene, two levels of selection need to be conducted. The first is to determine whether it is associated with any subtype at all. This is achieved using a bridge-type penalty. For a gene associated with at least one subtype, the second level of selection is to determine which subtype(s) it is associated with. This is achieved using a Lasso type penalty. The composition of the two penalties can achieve the desired two-level selection. Multiple SNPs may correspond to the same gene. The effect of gene j for subtype m is represented by the length- d_{jm} vector β_j^m . Here the penalty is imposed on the L_2 norm of β_j^m . Within a selected gene, no further SNP-level selection is conducted. If a gene is selected, all SNPs within this gene are selected.

3.2 Remarks

In the literature, the only method that can accommodate the same data settings and analysis goal is that in Ma et al. (2012), which uses thresholding for marker selection. Statistical properties of thresholding are extremely hard to establish. In addition, simulation in Section 4 shows that the proposed method outperforms the thresholding method. The existing penalization methods are not directly applicable here. The group bridge (Huang et al. 2009), group MCP (Liu et al. 2012), and other existing composite penalization methods conduct two-level selection. They have been applied to, for example, gene expression data, where the effect of a gene in a specific dataset is represented by a single regression coefficient. With SNP data, the effect of a gene with multiple SNPs is represented by a vector of regression coefficients. In addition, statistical properties of the existing group penalization methods have not been fully established for integrative analysis under the heterogeneity model. The tree-guided group Lasso (Kim and Xing 2010) and related methods are also applicable to multi-dataset settings. However, they assume that the outcomes have a natural tree structure, which is not present in this study. In addition, their statistical properties under the present

settings are not clear. As Lasso-type penalties are used, it is speculated that they may not have the selection consistency property.

An important gene may contain “noisy” SNPs. It is possible to extend the proposed method and conduct the third-level, within-gene-SNP selection. One possibility is to replace the L_2 norm with the L_1 norm (Lasso) or another penalty such as bridge or MCP. An advantage of this extension is that it can remove noises and further simplify the models. However, it may significantly increase computational complexity as the Lasso (or bridge, MCP) penalty function is not differentiable. In addition, it will make the theoretical development much more difficult. Identifying important genes (without further discriminating SNPs) can be sensible in practice. For example, in therapeutic development, it is easier to target a gene than a single SNP. Moreover, SNP data is “sparse” with at most three levels. Individual SNP effects are usually weak in cancer studies. Our unpublished numerical study suggests that to achieve satisfactory SNP-level selection, a much larger sample size is needed. If necessary, identifying important SNPs within genes can be conducted in downstream analysis after the identification of important genes. Here we acknowledge the possibility of within-gene selection but do not pursue it.

3.3 Computational algorithm

For subtype m , denote $\mathbf{Y}^m$ as the vector composed of $\sqrt{n}Y_{\omega}^m$'s, and $\mathbf{X}^m$ as the matrix composed of $\sqrt{n}\mathbf{X}_{\omega}^m$'s. Let $\mathbf{Y} = (\mathbf{Y}^{1'}, \dots, \mathbf{Y}^{M'})'$ and $\mathbf{X} = \text{diag}(\mathbf{X}^1, \dots, \mathbf{X}^M)$. Denote $\mathbf{X}_j$ as the submatrix of $\mathbf{X}$ corresponding to β_j . Then

$$L(\beta) = \sum_{m=1}^M \frac{1}{2} (\mathbf{Y}_{\omega}^m - \mathbf{X}_{\omega}^{m'} \beta^m)^2 = \frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^J \mathbf{X}_j \beta_j \right\|^2. \quad (2)$$

The penalized objective function is

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^J \mathbf{X}_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^J c_j \left(\sum_{m=1}^{M_j} \sqrt{d_{jm}} \|\beta_j^m\| \right)^{\gamma}. \quad (3)$$

Define

$$S(\beta, \theta) = \frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^J \mathbf{X}_j \beta_j \right\|^2 + \sum_{j=1}^J \theta_j^{1-1/\gamma} c_j^{1/\gamma} \sum_{m=1}^{M_j} \sqrt{d_{jm}} \|\beta_j^m\| + \tau_n \sum_{j=1}^J \theta_j, \quad (4)$$

where $\theta = (\theta_1, \dots, \theta_J)'$, and τ_n is a penalty parameter.

Proposition 3.1—If $\lambda_n = \tau_n^{1-\gamma} \gamma^{-\gamma} (1-\gamma)^{\gamma-1}$, then $\hat{\beta}$ minimizes the objective function in (3) if and only if $(\hat{\beta}, \hat{\theta})$ minimizes $S(\beta, \theta)$ subject to $\theta_j = 0$ for all j .

Proof is provided in Supplementary Materials. With $S(\boldsymbol{\beta}, \boldsymbol{\theta})$, optimization with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be conducted iteratively. With a fixed $\boldsymbol{\beta}$, optimization with respect to $\boldsymbol{\theta}$ has a simple analytic solution. With a fixed $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ minimizes a weighted-group-Lasso type objective function, which can be solved using existing algorithms. Motivated by such an observation, we propose the following algorithm.

- I. Denote $\boldsymbol{\beta}^{(0)}$ as the initial estimate. In our numerical study, the choice of initial estimate does not seem to have a big impact on the final estimate. More discussions are provided in Section 4. For simplicity, all components of $\boldsymbol{\beta}^{(0)}$ are set as 0.01. Set $s = 0$.
- II. $s = s + 1$. Compute

$$\theta_j^{(s)} = c_j \left(\frac{1-\gamma}{\gamma \tau_n} \right)^\gamma \left(\sum_{m=1}^{M_j} \sqrt{d_{jm}} \|\boldsymbol{\beta}_j^{m(s-1)}\| \right)^\gamma, \quad (5)$$

$$\boldsymbol{\beta}^{(s)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{2n} \|\mathbf{Y} - \sum_{j=1}^J \mathbf{X}_j \boldsymbol{\beta}_j\|^2 + \sum_{j=1}^J (\theta_j^{(s)})^{1-1/\gamma} c_j^{1/\gamma} \sum_{m=1}^{M_j} \sqrt{d_{jm}} \|\boldsymbol{\beta}_j^m\| \right), \quad (6)$$

- III. Repeat Step II until convergence

In the above algorithm, we use a fixed initial estimate. Using a “hot start” may be more efficient. $\boldsymbol{\beta}^{(s)}$ incurs the most computational cost and is obtained using a group coordinate descent algorithm (Huang et al. 2012), whose convergence can be derived following Tseng (2001). As in each iteration the nonnegative objective function decreases, the overall convergence can be guaranteed. However, as the group-bridge-type penalty is not convex, the algorithm may converge to a local minimizer depending on the initial value. For gene j , if it is not selected in Step s , then $\theta_j^{(s+1)}, \theta_j^{(s+2)}, \dots$ will all be zero, and this gene cannot be added back. Thus, over iterations, the selected gene sets are non-increasing. Properties of the proposed algorithm are also investigated numerically in Section 4. Research code written in R is available at works.bepress.com/shuangge/45/.

3.4 Tuning parameter selection

With bridge-type penalties, the value of γ is usually fixed. Theoretically speaking, different values of γ , as long as in the interval $(0, 1)$, lead to similar asymptotic results. In numerical studies, as $\gamma \rightarrow 1$, the bridge estimate behaves similarly to the Lasso estimate. On the other hand, as $\gamma \rightarrow 0$, it behaves similarly to the AIC/BIC penalized estimate. In simulation, we experiment with different γ values including 0.5 (which is the most commonly used in the literature), 0.7, and 0.9. The effect of λ_n is similar to that with other penalties. As $\lambda_n \rightarrow \infty$, fewer genes/SNPs are identified.

We use BIC for tuning parameter selection. Particularly, with a fixed λ , the optimal λ_n minimizes $\text{BIC}(\lambda_n) = \log \{ \|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\lambda_n)\|^2 / n \} + \log(n) \text{df}(\lambda_n) / n$. Here we use the notation $\hat{\boldsymbol{\beta}}(\lambda_n)$ to emphasize the dependence of $\hat{\boldsymbol{\beta}}$ on λ_n . An approximation to the degree of freedom is

proposed as $\tilde{df}(\lambda_n) = \sum_{j=1}^J \sum_{m=1}^{M_j} I(\|\hat{\beta}_j^m\| > 0) + \sum_{j=1}^J \sum_{m=1}^{M_j} \frac{\|\hat{\beta}_j^m\|}{\|\hat{\beta}_j^{mLS}\|} (d_{jm} - 1)$. Here $\hat{\beta}_j^{mLS}$ is obtained by fitting an AFT model (with least squares estimation) using the j th gene and m th subtype only. Proposition 3.1 suggests that, numerically, the proposed estimate can be transformed into a sequence of reweighted-group-Lasso estimates, whose tuning parameter selection and degree of freedom have been studied in Yuan and Lin (2006). The above proposal has been motivated by Yuan and Lin (2006) and others. We note that although intuitively reasonable, the validity of proposed tuning parameter selection has not been theoretically proved.

3.5 Practical considerations

With practical data, minor allele frequencies in some loci can be low. This may cause an instability problem in the Cholesky decomposition when some eigenvalues of the correlation matrices are small. In the proposed penalized selection, within-gene-SNP level selection is not conducted. To reduce the dimensionality within genes and to tackle the colinearity problem, we may first conduct principal component analysis (PCA) within genes. Specifically, we choose the number of PCs such that at least 90% of the total variation is explained. Then the PCs, as opposed to the original SNP measurements, are used in downstream analysis. This step can ensure that the smallest eigenvalues of the correlation matrices are not too small and that the Cholesky decomposition is stable.

4. Simulation

Three datasets, one for each cancer subtype, are simulated. Each dataset has 100 subjects. For each subject, the genotypes of 1,000 SNPs are simulated. They are first generated from multivariate normal distributions with marginal means 0 and standard deviations 1. Then the value of each SNP is set equal to 0, 1, or 2, depending on whether the continuous value is $< -c$, $\in [-c, c]$, or $> c$, where c is the 3rd quartile of the standard normal distribution. Genotypes j and k , if from different genes, have correlation coefficient $0.2^{|j-k|}$. For genotypes from the same genes, we consider the following two correlation structures. The first is the auto-regressive (AR) correlation, where genotypes j and k have correlation coefficient $\rho^{|j-k|}$. $\rho = 0.2, 0.5$, and 0.8 , corresponding to weak, moderate, and strong correlations, respectively. The second is the banded correlation structure. Here three scenarios are considered. Under the first scenario, genotypes j and k have correlation coefficient 0.2 if $|j - k| = 1$, 0.1 if $|j - k| = 2$, and 0 otherwise. Under the second scenario, genotypes j and k have correlation coefficient 0.5 if $|j - k| = 1$, 0.25 if $|j - k| = 2$, and 0 otherwise. Under the third scenario, genotypes j and k have correlation coefficient 0.6 if $|j - k| = 1$, 0.33 if $|j - k| = 2$, and 0 otherwise. In addition, consider the following cases of gene structures and nonzero regression coefficients.

Case 1 There are 200 genes, each with 5 SNPs. The nonzero regression coefficients for subtype 1 and 2 are (0.15, 0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1, 0.1), and the nonzero coefficients for subtype 3 are (0.1, 0.1, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1, 0.1). For each subtype, all SNPs within four genes have nonzero coefficients.

Case 2 There are 200 genes, each with 5 SNPs. The nonzero regression coefficients for subtype 1 and 2 are (0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1), and the nonzero coefficients for subtype 3 are (0.1, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1). For each subtype, four (out of five) SNPs within four genes have nonzero coefficients.

Case 3 There are 7 genes having 20 SNPs, 11 genes having 10 SNPs, 3 genes having 6 SNPs, 144 genes having 5 SNPs, and 3 genes having 4 SNPs. The nonzero regression coefficients for subtypes 1, 2 and 3 are ($\underbrace{0.1, \dots, 0.1}_{20}$, 0.2, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15). For each subtype, all SNPs within 1 gene with 20 SNPs, 1 gene with 4 SNPs, 1 gene with 5 SNPs, and 1 gene with 6 SNPs have nonzero coefficients.

Case 4 The gene structures are the same as Case 3. The nonzero regression coefficients for subtypes 1, 2 and 3 are ($\underbrace{0.1, \dots, 0.1}_{15}$, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.15, 0.15). For each subtype, 15 SNPs within 1 gene with 20 SNPs, 3 SNPs within 1 gene with 4 SNPs, 3 SNPs within 1 gene with 5 SNPs, and 5 SNPs within 1 gene with 6 SNPs have nonzero coefficients.

Case 5 The gene structures are the same as Case 3. The nonzero coefficients are equal to those in Case 3 plus normally distributed values with mean 0 and standard deviation 0.05.

Case 6 The gene structures are the same as Case 5, except that for the genes with 20 SNPs, only 5 SNPs within each gene have nonzero regression coefficients.

Under all cases, each subtype has 4 genes associated with prognosis. Under Case 1–4, the nonzero regression coefficients are fixed, and all important SNPs within the same gene have equal regression coefficients. Under Case 5–6, the nonzero regression coefficients are randomly generated. Under Case 1, 3, and 5, all SNPs within important genes are important. Under Case 2, 4, and 6, some SNPs within important genes are “noises”. Two heterogeneity models are considered. Under the first model, the three subtypes share three common susceptibility genes, and each subtype also has one unique susceptibility gene. The “unmatching rate” of susceptibility genes is 25%. Under the second model, the three subtypes share two common susceptibility genes, and each subtype has two unique susceptibility genes. The unmatching rate of susceptibility genes is 50%. As a special case of the heterogeneity model, the homogeneity model is also considered, under which all three subtypes have the same susceptibility genes.

The logarithms of event times are generated from the AFT models with intercept equal to 0.5 and normally distributed random errors. The logarithms of censoring times are generated as uniformly distributed and independent of the event times. The censoring distribution parameters are adjusted so that the overall censoring rate is about 45%.

We first more closely examine the computational algorithm. The algorithm contains an outer loop and an inner loop (in Step 2 which calls a group coordinate descent algorithm for reweighted group Lasso). For both loops, we use the L_2 -norm of the difference between two

consecutive estimates less than 0.001 as the convergence criterion. For both the outer and inner loops, and for all simulated datasets, convergence is achieved within 10 iterations. The proposed algorithm is computationally affordable. For example, under the simulation setting corresponding to Table 2, with fixed λ_n and γ , the analysis of one replicate takes 5 seconds on a regular laptop. The proposed algorithm requires an initial value. Beyond the one described in Section 3.3, we also experiment with generating the initial values from $N(0, 0.1)$ and $N(0, 0.01)$. Under the settings corresponding to Table 2 and 3, we compare estimates obtained using different initial values. With two settings and 100 replicates per setting, the mean of L_2 norms of differences (of estimates with different initial values) is 9.5×10^{-4} . It suggests that the proposed algorithm is relatively insensitive to initial values. With one simulated replicate, we show the L_2 norms of genes (across the three subtypes) as a function of the number of iterations in Web Figure 1 (Supplementary Materials). It is clear that the sets of selected genes are non-increasing.

Beyond the proposed method, simulated data are also analyzed using a thresholding method (Ma et al. 2012, referred to as “Threshold” below and in the tables) and a meta-analysis method. With Threshold, Ma et al. (2012) proposes selecting the tuning parameters using 5-fold cross validation. This method is referred to as “Threshold-CV”. In addition, to make a fair comparison, we also consider “Threshold-BIC”, which uses the BIC criterion to select tunings. With the meta-analysis method, each subtype is analyzed using the group Lasso (GLasso), where a group corresponds to one gene with multiple SNPs. Then the identified gene lists are combined across subtypes. The GLasso has been adopted in a large number of studies and serves as benchmark here. With the proposed method, we experiment with $\gamma = 0.5, 0.7$, and 0.9 . Summary statistics based on 100 replicates are shown in Table 2–3 and Web Table 1–12 (Supplementary Materials).

Simulation suggests that performance of the proposed method depends on the value of γ . As γ increases, in general, more true positives and more false positives are identified. This result is reasonable considering that $\gamma \rightarrow 0$ leads to an AIC/BIC type penalty, and $\gamma \rightarrow 1$ leads to a Lasso type penalty. Our limited simulation suggests better performance with a moderate to large γ value. Performance depends on correlation structure. In general, as correlation gets stronger, more true positives can be identified. Performance also depends on the unmatching rate. Under the homogeneity model (unmatching rate=0), the distinctions between signals and noises are stronger, making important genes easier to be identified. Under all simulation settings, when taking both the number of true positives and model size into consideration, we observe that the proposed method outperforms the two alternatives. For example in Table 2, under the AR correlation with $\rho = 0.5$, Threshold-CV identifies 11.1 true positives with a model size of 33.6; Threshold-BIC identifies 10.6 true positives with a model size of 28.2; GLasso identifies 6.7 true positives with a model size of 39.7; and the proposed method identifies 10.7, 10.8, and 11.0 true positives, with model sizes of 14.4, 14.7, and 17.1, respectively, under different γ values. We have experimented with a few other settings and reached similar conclusions.

5. Analysis of NHL Genetic Association Data

NHL is the fifth leading cause of cancer incidence and mortality in the US and remains poorly understood and largely incurable. It has multiple subtypes and is highly diverse. For example, DLBCL (the largest subtype) is aggressive, whereas FL (the second largest subtype) is indolent. Chromosomal translocations such as t(3, 22) are specific to DLBCL, whereas others such as t(14, 18) are specific to FL. On the other hand, different subtypes share common susceptibility genes/SNPs. Genes in the cell cycle, multiple signaling, RAS, and DNA repair pathways are involved in the development and progression of multiple cancers including NHL. For NHL specifically, Han et al. (2010) and Ma et al. (2010) find that SNPs in multiple genes, such as BRCA2, CASP3, IRF1, BCL2, NAT2, and ALXO12B, are associated with both DLBCL and FL.

We conducted a genetic association study, searching for SNPs/genes associated with the overall survival of NHL patients. The prognostic cohort consisted of 575 NHL patients, among whom 496 donated either blood or buccal cell samples. All cases were classified into NHL subtypes according to the World Health Organization classification system. Specifically, 155 had DLBCL, 117 had FL, 57 had CLL/SLL, 34 had MZBL, 37 had T/NK-cell lymphoma, and 96 had other subtypes. Because of sample size consideration, we focus on DLBCL, FL, and CLL/SLL, the three largest subtypes in this dataset. The study cohort was assembled in Connecticut between 1996 and 2000. Vital status of all subjects was abstracted from the CTR (Connecticut Tumor Registry) in 2008.

A candidate gene approach was taken in genotyping. Specifically, a total of 1462 tag SNPs from 210 candidate genes related to immune response were genotyped using a custom-designed GoldenGate assay. In addition, 302 SNPs in 143 candidate genes previously genotyped by Taqman assay were also included. There were a total of 1764 SNPs, representing 333 genes. Data preprocessing is conducted. Subjects with more than 20% SNPs missing are removed from analysis. Then SNPs with more than 20% missing are removed. Genotyping data were missing for the following reasons: the amount of DNA was too low, samples failed to amplify, samples amplified but their genotypes could not be determined due to ambiguous results, or DNA quality was poor. The remaining missing SNP measurements are imputed. A total of 1,633 SNPs pass processing, representing 238 genes.

For DLBCL, 139 patients pass processing. Among them, 61 died, with survival times ranging from 0.47 to 10.46 years (mean 4.16 years). For the 78 censored patients, the follow-up times range from 5.58 to 11.45 years (mean 9.08 years). For FL, 102 patients pass processing. Among them, 33 died, with survival times ranging from 0.91 to 10.23 years. For the 69 censored patients, the follow-up times range from 4.96 to 11.39 years, with mean 8.83 years. For CLL/SLL, 50 patients pass processing. Among them, 27 died, with survival times ranging from 1.91 to 10.13 years (mean 4.85 years). For the 23 censored patients, the follow-up times range from 4.92 to 11.07 years, with mean 8.83 years.

We analyze data using the proposed method. Table 4 contains the L_2 norms of identified genes. Web Table 13–15 (Supplementary Materials) contain the estimated regression coefficients for SNPs. Fourteen genes are identified as associated with the overall survival of

DLBCL. Twelve genes are identified as associated with FL. And five genes are identified as associated with CLL/SLL. Among the identified genes, MBP and STAT4 are shared by all three subtypes, ALOX5, IL10, IRAK2, LMAN1, MIF, and NCF4 are shared by two subtypes, and thirteen other genes are identified as subtype-specific. We search published literature and find that the identified genes may have important implications. More details are provided in Web Appendix A.

The relative stability of identified genes is evaluated using a random sampling approach (Huang and Ma 2010). In particular, we randomly sample 3/4 of the subjects without replacement and apply the proposed method. This process is repeated 100 times. For each gene, we compute the probability of it being identified out of the 100 samplings. This probability is referred to as the observed occurrence index (OOI) and measures the relative stability. Table 4 shows that only gene IL10 for DLBCL has a low occurrence index. All other observed occurrence indexes are high, suggesting satisfactory stability. We also adopt the approach in Huang and Ma (2010) and evaluate prediction. In particular, genes are identified, and models are constructed using the randomly sampled subjects. Then prediction is made for the remaining subjects (testing). Based on the predicted $\mathbf{X}^{m'}\hat{\boldsymbol{\beta}}^n$ s, the testing subjects are separated into two risk groups. The logrank statistic is computed and evaluates whether the model can separate subjects into groups with significantly different survival. Using the logrank for prediction evaluation has been adopted in quite a few publications (see for example Huang and Ma 2010 and others). It can be more straightforward for censored survival data than some alternatives such as the prediction error and more easily to compute than for example the Receiver Operating Characteristic-based measures. The above process is repeated 100 times, and the mean logrank statistic is computed as 7.1 (p -value 0.0077), suggesting satisfactory prediction.

Data are also analyzed using multiple alternative methods. Ma et al. (2012) applies the Threshold-CV method and identifies 12 (DLBCL), 11 (FL), and 15 (CLL/SLL) genes, respectively. One gene is shared by all three subtypes, and 6 genes are shared by two subtypes. Prediction evaluation generates a logrank statistic of 4.417 (p -value 0.036). Using BIC for tuning selection, the thresholding method identifies 13 (DLBCL), 9 (FL), and 12 (CLL/SLL) genes, respectively. Four genes are shared by two subtypes. The prediction logrank statistic is 4.132 (p -value 0.042). The GLasso method identifies 26 (DLBCL), 17 (FL), and 8 (CLL/SLL) genes, respectively. Some additional results are provided in Supplementary Materials. There are 3 genes shared by two subtypes, and all others are subtype-specific. The prediction logrank statistic is 0.2 (p -value 0.65). We also consider the following approach. The first PC of each gene is extracted and used for analysis. Thus the group size is 1. The proposed method then simplifies to the group Bridge, and the GLasso method simplifies to a Lasso method. The estimation results are shown in Web Table 19 and 20–22. In Web Table 19, 22 (DLBCL), 14 (FL), and 15 (CLL/SLL) genes are identified, respectively. Among them, 6 are shared by all subtypes, 11 are shared by two subtypes, and 8 are subtype-specific. The prediction logrank statistic is 5.8 (p -value 0.016). In Web Table 20–22, 67 (DLBCL), 30 (FL), and 28 (CLL/SLL) genes are identified, respectively. Among them, 10 genes are shared by two subtypes. The prediction logrank statistic is 6.0 (p -value

0.015). The proposed method identifies genes different from the alternatives and has the best prediction performance.

6. Discussion

With prognosis data on multiple subtypes of the same cancer, we have developed a penalized integrative analysis method that can identify important genes containing SNPs associated with multiple subtypes and allow for subtype-specific susceptibility genes. The proposed method is realized using an effective iterative algorithm. Under mild conditions, it has the selection consistency properties. Simulation shows that it outperforms the thresholding and GLasso-based meta-analysis methods. In the analysis of NHL prognosis data, it identifies multiple genes shared by two or three subtypes as well as subtype-specific genes. The shared genes have important biological implications. The proposed method also leads to better prediction performance.

In simulation and data analysis, we have focused on the scenario with multiple subtypes of the same cancer and the “SNP-within-gene” structure. The proposed method is directly applicable to the analysis of multiple datasets on the same prognosis outcome and the analysis of multiple datasets on different types of diseases. In addition, it is also applicable to the “gene-with-pathway (statistical cluster)” structure. With minor modifications, the proposed penalization method can be used to analyze prognosis data under other models and etiology data. In data analysis, our preliminary search shows that the genes shared by multiple NHL subtypes have important implications. However, because of the following limitations, the analysis results should be interpreted with caution. First, the sample size is limited. Second, the NHL study took a candidate gene approach. It is possible that important genes have been missed in the profiling stage. Third, the proposed evaluation is cross-validation based. Although it compares different methods on the same ground, it does not use completely independent data. More and independent studies are needed to fully comprehend the data analysis results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor, associate editor, and two reviewers for very careful review and insightful comments. This study was supported by CA142774, CA165923, and CA152301 from National Institute of Health, DMS1208225 from National Science Foundation, 2012LD001 from National Bureau of Statistics of China, and 13CTJ001 from National Social Science Foundation of China.

References

- Goh KI, Choi IG. Exploring the human diseasome: the human disease network. *Brief Funct Genomics*. 2012; 11(6):533–542. [PubMed: 23063808]
- Han X, Li Y, Huang J, Zhang Y, Holford T, Lan Q, et al. Identification of predictive pathways for non-Hodgkin lymphoma prognosis. *Cancer Informatics*. 2010; 9:281–292. [PubMed: 21245948]
- Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*. 2010; 16:176–195. [PubMed: 20013308]

- Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. *Biometrika*. 2009; 96(2):339–355. [PubMed: 20037673]
- Huang J, Wei F, Ma S. Semiparametric regression pursuit. *Statistica Sinica*. 2012; 22:1403–1426. [PubMed: 23559831]
- Kim, S.; Xing, EP. Tree-guided group Lasso for multi-task regression with structured sparsity. *Proceedings of the 27th International Conference on Machine Learning*; 2010.
- Liu J, Huang J, Ma S. Integrative analysis of cancer diagnosis studies with composite penalization. *Scandinavian Journal of Statistics*. 2012 In press.
- Ma S, Huang J, Moran M. Identification of genes associated with multiple cancers via integrative analysis. *BMC Genomics*. 2009; 10:535. [PubMed: 19919702]
- Ma S, Zhang Y, Huang J, Han X, Holford T, Lan Q, et al. Identification of Non-Hodgkin's lymphoma prognosis signatures using the CTGDR method. *Bioinformatics*. 2010; 26:15–21. [PubMed: 19850755]
- Ma S, Zhang Y, Huang J, Huang Y, Lan Q, Rothman N, et al. Integrative analysis of cancer prognosis data with multiple subtypes using regularized gradient descent. *Genetic Epidemiology*. 2012; 36:829–838. [PubMed: 22851516]
- Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *PNAS*. 2004; 101(25):9309–9314. [PubMed: 15184677]
- Stute W. Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*. 1996; 23:461–471.
- Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optimization Theory and Applications*. 2001; 109:475–494.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *JRSSB*. 2006; 68:49–67.
- Zhang Y, Dai Y, Zheng T, Ma S. Risk factors of Non-Hodgkin lymphoma. *Expert Opinion on Medical Diagnostics*. 2011; 5:539–550. [PubMed: 22059093]
- Zhang Y, Lan Q, Rothman N, Zhu Y, Zahm SH, Wang SS, et al. A putative exonic splicing polymorphism in the BCL6 gene and the risk of non-Hodgkin lymphoma. *J Natl Cancer Inst*. 2005; 97:1616–1618. [PubMed: 16264183]

Table 1

Regression coefficients for a study with three subtypes, four genes, and eight SNPs. An empty cell corresponds to a zero regression coefficient.

Gene	SNP	Subtype		
		S1	S2	S3
1	1_1	0.20	0.19	0.21
	1_2	-0.22	-0.19	-0.21
2	2_1	0.18	0.21	
	2_2	-0.21	-0.21	
3	3_1			0.21
	3_2			-0.18
4	4_1			
	4_2			

Table 2

Simulation under the heterogeneity model: unmatching rate=25% and nonzero regression coefficients under Case 1. In each cell, the first row is the number of true positives (standard deviation), and the second row is the model size (standard deviation).

Correlation	Threshold-CV	Threshold-BIC	GLasso	Proposed		
				$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
AR $\rho = 0.2$	6.9(5.5)	6.2(5.4)	5.7(2.4)	6.7(5.2)	7.4(4.6)	8.9(2.8)
	25.8(11.7)	22.9(10.5)	36.4(16.5)	9.5(7.6)	10.7(6.6)	20.1(6.3)
AR $\rho = 0.5$	11.1(3.6)	10.6(3.6)	6.7(2.4)	10.7(3.0)	10.8(2.8)	11.0(1.9)
	33.6(20.1)	28.2(18.3)	39.7(19.3)	14.4(4.5)	14.7(4.0)	17.1(4.5)
AR $\rho = 0.8$	10.7(3.4)	10.9(3.3)	9.4(2.3)	12.0(0.2)	12.0(0.2)	11.9(0.2)
	31.9(10.6)	26.7(9.9)	50.2(19.3)	14.4(1.9)	14.6(1.9)	15.6(2.7)
Banded 1	8.3(4.0)	8.5(3.5)	5.3(2.6)	7.8(5.0)	8.6(4.4)	8.9(3.6)
	33.1(7.8)	30.7(7.6)	33.8(16.2)	11.0(7.1)	11.9(5.9)	20.5(7.0)
Banded 2	9.4(3.5)	8.8(3.5)	7.5(2.8)	10.6(3.3)	11.2(2.2)	11.3(1.7)
	36.7(19.3)	32.8(19.8)	45.5(21.7)	14.6(4.8)	15.5(3.7)	21.1(7.0)
Banded 3	10.7(3.2)	10.2(3.1)	7.9(2.4)	11.5(1.9)	11.5(1.5)	11.6(1.1)
	35.4(16.3)	32.1(15.4)	44.2(17.7)	15.4(3.0)	15.4(2.7)	18.8(5.2)

Table 3

Simulation under the heterogeneity model: unmatched rate=50% and nonzero regression coefficients under Case 1. In each cell, the first row is the number of true positives (standard deviation), and the second row is the model size (standard deviation).

Correlation	Threshold-CV	Threshold-BIC	GLasso	Proposed		
				$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$
AR $\rho = 0.2$	6.2(3.4)	6.4(3.2)	4.6(2.3)	3.7(4.1)	5.8(3.8)	8.3(2.1)
	24.7(18.2)	22.1(17.3)	30.7(16.9)	5.4(7.0)	10.3(6.9)	22.2(6.8)
AR $\rho = 0.5$	8.9(3.7)	8.6(3.3)	6.5(3.0)	9.6(3.5)	9.7(3.1)	10.1(2.3)
	27.5(20.0)	25.5(19.6)	36.1(20.1)	16.8(7.5)	17.2(6.9)	21.1(6.8)
AR $\rho = 0.8$	10.8(2.4)	11.0(2.2)	9.7(2.2)	11.5(0.9)	11.5(0.7)	11.6(0.7)
	49.2(16.8)	44.4(15.9)	54.5(17.5)	19.4(3.9)	18.4(4.1)	20.2(5.8)
Banded 1	5.3(3.3)	5.2(3.3)	4.7(2.4)	4.4(4.1)	5.9(3.8)	7.7(2.9)
	33.6(16.7)	30.1(15.7)	34.5(17.6)	6.7(7.2)	10.8(7.7)	20.8(9.2)
Banded 2	9.2(4.1)	9.9(3.9)	7.5(2.5)	8.4(3.9)	9.1(3.1)	10.0(1.7)
	41.4(18.3)	39.8(18.0)	47.1(19.4)	14.4(7.7)	15.9(6.1)	23.5(7.1)
Banded 3	8.8(2.5)	8.7(2.5)	7.3(2.4)	9.5(2.9)	9.9(2.6)	10.1(2.2)
	30.9(19.2)	26.5(18.9)	43.2(19.6)	16.6(6.7)	17.7(5.9)	22.9(7.6)

Table 4

Analysis of the NHL data using the proposed method. For a gene, the L_2 -norm of estimate and OOI (observed occurrence index).

Gene	DUBCL		FL		CLL/SLL	
	L_2 -norm	OOI	L_2 -norm	OOI	L_2 -norm	OOI
ALOX12			0.02	0.83		
ALOX15B	0.01	0.76				
ALOX5	0.02	0.71			0.01	0.64
CLCA1			0.02	0.83		
CSF2	0.02	0.86				
DEFB1	0.03	0.97				
IL10	1.E-04	0.21			0.02	0.62
IL17C			0.02	0.78		
IRAK2			0.02	0.88	0.01	0.80
LIG4			0.01	0.87		
LMAN1	0.02	0.71	0.02	0.77		
MBP	0.01	0.68	4.E-03	0.60	0.04	0.68
MCP	0.01	0.83				
MEFV	0.02	0.85				
MIF	0.02	0.83	1.E-03	0.53		
MUC6	0.03	0.99				
NCF4	0.01	0.64	0.01	0.64		
PTK9L			0.01	0.66		
SERPINB3	0.01	0.55				
SOD3			4.E-03	0.47		
STAT4	0.02	0.95	0.01	0.88	0.01	0.91