

Identification of Gene-environment Interactions in Cancer Prognosis Studies Using Penalization

Jin Liu¹, Jian Huang², Yawei Zhang¹, Qing Lan³, Nathaniel Rothman³,
Tongzhang Zheng¹, and Shuangge Ma¹

¹School of Public Health, Yale University

²Department of Statistics & Actuarial Science, Department of Biostatistics,
University of Iowa

³Division of Cancer Epidemiology and Genetics, National Cancer Institute,
NIH

September 3, 2012

Abstract

High-throughput cancer studies have been extensively conducted, searching for genetic risk factors independently associated with prognosis beyond clinical and environmental risk factors. Many studies have shown that the gene-environment interactions may have important implications. Some of the existing methods, such as the commonly adopted single-marker analysis, may be limited in that they cannot accommodate the joint effects of a large number of genetic markers or use ineffective marker identification techniques. In this study, we analyze cancer prognosis studies, and adopt the AFT (accelerated failure time) model to describe survival. A weighted least squares approach, which has the lowest computational cost, is adopted for estimation. For the identification of $G \times E$ interactions and main effects, we adopt a group sparse penalization approach, which has an intuitive formulation, can accommodate the joint effects of a large number of markers, and is computationally affordable. Simulation study shows satisfactory performance of the penalization approach. Analysis of an NHL (non-Hodgkin lymphoma) prognosis study with SNP measurements shows that the proposed approach may identify markers with important implications and satisfactory prediction performance and reproducibility. Analysis of a follicular lymphoma gene expression study is also conducted.

Keywords: Gene-environment interaction; Cancer Prognosis; Marker selection; Penalization.

1 Introduction

In cancer research, high-throughput profiling has been extensively conducted, searching for genetic markers that may be independently associated with prognosis beyond clinical and environmental risk factors. Many studies, particularly the early ones, have been focused on studying the additive effects of risk factors. On the other hand, studies have also shown that the interactions between genetic and clinical/environmental risk factors (also referred to as G×E interactions) may have important implications beyond the additive, main effects. For comprehensive discussions of the existing methods, we refer to Hunter (2005), North and Martin (2008), Thomas (2010), Amato et al. (2010) and others. In the literature, multiple families of methods have been developed, including for example the joint method and stratification method. In this study, we focus on the statistical modeling method, where the interactions are described using products of variables in statistical models.

Consider a cancer prognosis study with n samples. Denote T as the survival time of interest. Denote $Z = (Z_1, \dots, Z_p)$ as the p SNPs measured on each subject, and $X = (X_1, \dots, X_q)$ as the q clinical/environmental risk factors. A popular, model-based approach proceeds as follows. (1) For $j = 1, \dots, p$, fit the regression model $T \sim \phi(\sum_{k=1}^q \alpha_k X_k + \gamma_j Z_j + \sum_{k=1}^q \beta_{kj} X_k Z_j)$, where ϕ is the known link function, and α_k s, γ_j , and β_{kj} s are the unknown regression coefficients. As usually $q \ll n$, for each j , this step can be carried out using standard techniques and software. Denote p_{kj} as the p-value of $\hat{\beta}_{kj}$, the estimate of β_{kj} ; (2) With $\{p_{kj} : k = 1, \dots, q, j = 1, \dots, p\}$, conduct multiple comparison adjustment. Approaches such as the FDR (false discovery rate) can be adopted to identify significant interaction effects. Multiple approaches reviewed in Hunter (2005) and Thomas (2010) belong to this category. Different approaches may differ in terms of statistical models, hypothesis testing methods, and multiple comparison adjustment techniques, however, share the common strat-

egy of studying one genetic marker at a time. There are also nonparametric, more robust approaches, for example the well-known MDR (multifactor dimensionality reduction; Moore et al. 2006), that shares a similar single-marker strategy. The most significant advantage of such approaches is computational simplicity. As step (1) only involves low-dimensional models and can be conducted in a parallel manner, the overall computational cost is low.

The development and progression of complex diseases such as cancer are associated with the combined effects of multiple clinical, environmental, and genetic risk factors and their interactions. Unless under strong orthogonality conditions (which are unlikely to hold in practice), estimation and inference based on the marginal models can be biased. With additive genetic effects, the discrepancy between marginal and joint analysis has been discussed in Witten and Tibshirani (2010), Zhang et al. (2011) and others.

In this article, we study $G \times E$ interactions in high-throughput cancer prognosis studies. We adopt a model-based approach, where the detection of interactions amounts to conducting estimation and inference with β_{kj} s. Unlike in some of the existing studies, we consider the joint effects of a large number of genetic markers ($p \gg n$) and their interactions with clinical/environmental risk factors. Here the main challenge is to properly accommodate the high dimensionality.

We adopt an AFT (accelerated failure time) model to describe cancer survival. Compared with alternatives such as the Cox model, the AFT model has a simpler form and lower computational cost, which are especially desirable with high-throughput data. In addition, the estimated regression coefficients may have more lucid interpretations. For detecting interactions, we adopt penalization, with which identifying important interactions amounts to finding those terms with nonzero estimated regression coefficients. The advantages of penalization approaches with high-throughput data have been well established elsewhere and will not be reiterated here. For the present problem, the most significant advantage of

penalization is that it can accommodate the joint effects of a large number of markers with affordable computational cost.

2 Data and Model

With slight abuse of notation, denote T as the logarithm (or another known monotone transformation) of the failure time. The AFT model assume that

$$\begin{aligned} T &= \alpha_0 + \sum_{k=1}^q \alpha_k X_k + \sum_{j=1}^p \left(\gamma_j Z_j + \sum_{k=1}^q \beta_{kj} X_k Z_j \right) + \epsilon \\ &= \alpha_0 + \sum_{k=1}^q \alpha_k X_k + \sum_{j=1}^p b'_j W_j + \epsilon. \end{aligned} \quad (1)$$

Here α_0 is the unknown intercept, $b_j = (\gamma_j, \beta_{1j}, \dots, \beta_{qj})'$, $W_j = (Z_j, X_1 Z_j, \dots, X_q Z_j)'$, and ϵ is the random error. In (1), W_j and b_j represent all effects (including main effects and interactions) corresponding to the j th SNP. Under right censoring, for each subject we observe (Y, δ, X, Z) , where $Y = \min(T, C)$, C is the logarithm of the censoring time, and $\delta = I\{T \leq C\}$ is the event indicator.

When the distribution of ϵ is known, the parametric likelihood function can be easily constructed. Here we consider the more flexible case where this distribution is unknown. In the literature, multiple estimation approaches have been developed, including for example the Buckley-James and rank-based approaches. In this study, we adopt the weighted least squares estimator (Stute 1996), which to the best of our knowledge, has the lowest computational cost. This property is especially desirable with high-throughput data.

Assume n iid observations. We use the subscript “ i ” to denote the i th subject. Let \hat{F} be the Kaplan-Meier estimator of the distribution function F of T . \hat{F} can be written as $\hat{F}(y) = \sum_{i=1}^n \omega_i I\{Y_{(i)} \leq y\}$, where ω_i s are the jumps in the Kaplan-Meier estimator and can

be computed as

$$\omega_1 = \frac{\delta_{(1)}}{n}, \quad \omega_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \quad i = 2, \dots, n.$$

ω_i s are also referred to as the Kaplan-Meier weights. Here $Y_{(1)} \leq \dots \leq Y_{(n)}$ are the order statistics of Y_i 's, and $\delta_{(1)}, \dots, \delta_{(n)}$ are the associated censoring indicators. Denote $(X_{(i)}, Z_{(i)})$ as the environmental and genetic measurements associated with $(Y_{(i)}, \delta_{(i)})$.

Denote $\alpha = (\alpha_1, \dots, \alpha_q)'$, $b = (b'_1, \dots, b'_p)'$, and $W = (W'_1, \dots, W'_p)'$. Stute (1996) proposed the weighted least squares estimator $(\hat{\alpha}_0, \hat{\alpha}, \hat{b})$ that minimizes

$$\frac{1}{2n} \sum_{i=1}^n \omega_i (Y_{(i)} - \alpha_0 - \alpha' X_{(i)} - b' W_{(i)})^2. \quad (2)$$

We center $X_{(i)}$, $W_{(i)}$ and $Y_{(i)}$ using their ω_i -weighted means, respectively. Define

$$\bar{X}_w = \sum_{i=1}^n \omega_i X_{(i)} / \sum_{i=1}^n \omega_i, \quad \bar{W}_w = \sum_{i=1}^n \omega_i W_{(i)} / \sum_{i=1}^n \omega_i, \quad \bar{Y}_w = \sum_{i=1}^n \omega_i Y_{(i)} / \sum_{i=1}^n \omega_i.$$

Let $X_{\omega(i)} = \sqrt{\omega_i} (X_{(i)} - \bar{X}_w)$, $W_{\omega(i)} = \sqrt{\omega_i} (W_{(i)} - \bar{W}_w)$ and $Y_{\omega(i)} = \sqrt{\omega_i} (Y_{(i)} - \bar{Y}_w)$, respectively.

With the weighted centered values, the intercept is zero. The weighted least squares objective function can be written as

$$L(\alpha, b) = \frac{1}{2n} \sum_{i=1}^n (Y_{\omega(i)} - \alpha' X_{\omega(i)} - b' W_{\omega(i)})^2. \quad (3)$$

Denote $Y = (Y_{\omega(1)}, \dots, Y_{\omega(n)})'$, $X = (X_{\omega(1)}, \dots, X_{\omega(n)})'$ and $W = (W_{\omega(1)}, \dots, W_{\omega(n)})'$. Then $L(\alpha, b) = \frac{1}{2n} \|Y - X\alpha - Wb\|^2$, where $\|\cdot\|$ is the ℓ_2 norm.

3 Identification of G \times E Interactions

We adopt penalization for the identification of main genetic effects and interactions that are associated with prognosis. With penalization, marker selection and regularized estimation

are achieved simultaneously. Consider the penalized estimate

$$(\hat{\alpha}, \hat{b}) = \operatorname{argmin} \left\{ L(\alpha, b) + \sum_{j=1}^p \rho(\|b_j\|; \sqrt{d}\lambda_1, \xi) + \sum_{j=1}^p \sum_{k=2}^{q+1} \rho(|b_{jk}|; \lambda_2, \xi) \right\}. \quad (4)$$

In the above formulation, $\rho(t; \lambda, \xi) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda\xi}\right)_+ dx$ is the MCP penalty (Zhang 2010). $d = q + 1$ is the “size” of b_j , and can be absorbed into λ_1 . We keep the present notation to be consistent with the existing group penalization methods. λ_1 and λ_2 are the data-dependent tuning parameters. ξ is the regularization parameter. b_{jk} is the k th element of b_j .

Formulation (4) has been motivated by the following considerations. In our analysis, SNPs are the functional units. The penalty is the sum over p terms, with one for each SNP. For the j th SNP, its effect is represented with b_j , a length $q + 1$ vector. With the first penalty in (4), we determine whether $b_j \equiv 0$, that is, whether the j th SNP has any effect at all. This step is achieved using the group MCP (gMCP) penalty. The gMCP has been proposed in Huang et al. (2012), although under dramatically different data and model settings, and shown to have performance better than for example group Lasso. If $b_j \neq 0$, then either the main effect or interaction or both are nonzero. In the analysis of interaction, a common practice is that if the interaction is important, then the main effect is kept even if it is not important. Thus, in the second penalty, we only penalize the interaction terms and determine which are nonzero. This step amounts to examining q individual interaction terms, and is achieved using the MCP penalty. The sum of the two penalties can thus identify important SNPs as well as important interaction terms. As the dimensionality of clinical and environmental risk factors is usually low, we do not conduct selection with such terms. When it is of interest to discriminate important environmental risk factors from unimportant ones, penalization on α can be imposed.

Formulation (4) shares a similar spirit with Friedman et al. (2010), which also proposes using the sum of group and individual penalties. This study differs from Friedman et al.

(2010) in that one “group” corresponds to one SNP and its interactions, as opposed to a group of multiple variables. Second, to respect the “main effects, interaction” hierarchy, the individual penalty is only imposed on the interactions. Third, we replace the Lasso-type penalties with MCP penalties, which have better selection properties. Last, we analyze $G \times E$ interactions in cancer prognosis studies, where the data and model settings are considerably more complex than those in Friedman et al. (2010). In a recent study, Bien et al. (2012) proposes Lasso-type penalization to study interaction. Unlike in Bien et al. (2012), we focus on $G \times E$ interactions and thus can analyze a much larger number of markers. In addition, we only require that if $\|(\beta_{1j}, \dots, \beta_{qj})\| \neq 0$ then $\gamma_j \neq 0$, but do not impose restriction on their magnitudes, which may significantly reduce computational complexity. In addition, our data analysis suggests that the magnitude restriction in Bien et al. (2012) is not necessarily true.

3.1 Computation

For computation, we consider the following iterative algorithm: (i) Initialize $\hat{\alpha} = 0$ and $\hat{b} = 0$ component-wise; (ii) Compute $\hat{\alpha} = (X'X)^{-1}X'(Y - W\hat{b})$; (iii) Compute \hat{b} as the minimizer of (4) with α fixed at $\hat{\alpha}$; (iv) Iterate steps (ii) and (iii) until convergence. In this algorithm, the most challenging step is (iii), for which we adopt a group coordinate descent (GCD) algorithm. The GCD algorithm optimizes the objective function with respect to one “group” of regression coefficients at a time (which correspond to the main and interaction effects of one SNP), and iteratively cycles through all groups. The overall cycling is repeated multiple times until convergence.

More specifically, the GCD algorithm proceeds as follows. For $j = 1, \dots, p$, given the parameter vectors b_l ($l \neq j$) fixed at their current estimates \hat{b}_l , we seek to minimize the objective function defined in (4) with respect to the j th group parameter b_j . Here only the terms involving b_j matter. Formulation (4) as a function of b_j while keeping all other

parameter vectors fixed is

$$R(b_j) = \frac{1}{2n} \|r_{b_{-j}} - W_j b_j\|^2 + \rho(\|b_j\|; \sqrt{d}\lambda_1, \xi) + \sum_{k=2}^{q+1} \rho(|b_{jk}|; \lambda_2, \xi). \quad (5)$$

Here W_j is the component of W corresponding to b_j , $r_{b_{-j}} = Y - X\hat{\alpha} - W_{-j}\hat{b}_{-j}$, and W_{-j} and \hat{b}_{-j} are W and \hat{b} with components corresponding to the j th SNP removed. Without loss of generality, assume that W_j is orthogonalized group-wise such that $\frac{1}{n}W_j'W_j = I$ for $j = 1, \dots, p$. As discussed in Huang et al. (2012), this normalization can be achieved by changing $\|b_j\|$ to $\|b_j\|_{W_j}$.

The first order derivative of (5)

$$\frac{\partial R(b_j)}{\partial b_j} = -\frac{1}{n}W_j'r_{b_{-j}} + b_j + \frac{b_j}{\|b_j\|} \begin{cases} \sqrt{d}\lambda_1 - \frac{\|b_j\|}{\xi}, & \text{if } \|b_j\| \leq \xi\sqrt{d}\lambda_1 \\ 0, & \text{if } \|b_j\| > \xi\sqrt{d}\lambda_1 \end{cases} + t \quad (6)$$

where

$$t = \left(0, \text{sgn}(b_{j2}) \begin{cases} \lambda_2 - \frac{|b_{j2}|}{\xi}, & \text{if } |b_{j2}| \leq \xi\lambda_2 \\ 0, & \text{if } |b_{j2}| > \xi\lambda_2 \end{cases}, \dots, \text{sgn}(b_{jq+1}) \begin{cases} \lambda_2 - \frac{|b_{jq+1}|}{\xi}, & \text{if } |b_{jq+1}| \leq \xi\lambda_2 \\ 0, & \text{if } |b_{jq+1}| > \xi\lambda_2 \end{cases} \right)'$$

and $\text{sgn}(a) = 1, 0, -1$ for $a > 0, = 0, < 0$, respectively.

By setting expression (6) equal to zero, we have:

$$-u_j + g(b_j)b_j + t = 0 \quad (7)$$

where $u_j = \frac{1}{n}W_j'r_{b_{-j}}$ and $g(b_j) = \left(1 + \frac{1}{\|b_j\|} \begin{cases} \sqrt{d}\lambda_1 - \frac{\|b_j\|}{\xi}, & \text{if } \|b_j\| \leq \xi\sqrt{d}\lambda_1 \\ 0, & \text{if } \|b_j\| > \xi\sqrt{d}\lambda_1 \end{cases} \right)$.

Denote u_{jk} as the k th element of u_j . First, in $g(b_j)$ fix b_j at the current estimate \hat{b}_j . We use g short for $g(\hat{b}_j)$. It is straightforward to see that the solution to equation (7) is $\widehat{gb}_{j1} = u_{j1}$ and $\widehat{gb}_{jk} = \begin{cases} \frac{S_1(u_{jk}, \lambda_2)}{1 - \frac{1}{\xi g}}, & \text{if } |u_{jk}| \leq \xi\lambda_2 g \\ u_{jk}, & \text{if } |u_{jk}| > \xi\lambda_2 g \end{cases}$, for $k = 2, \dots, q+1$, where $S_1(u, \lambda) = \text{sgn}(u)(|u| - \lambda)_+$. For $k = 1, \dots, q+1$, we set $v_{jk} = \widehat{gb}_{jk}$ and $v_j = (v_{j1}, \dots, v_{jq+1})'$. Taking v_j back into its definition, we have

$$b_j + \frac{b_j}{\|b_j\|} \begin{cases} \sqrt{d}\lambda_1 - \frac{\|b_j\|}{\xi}, & \text{if } \|b_j\| \leq \xi\sqrt{d}\lambda_1 \\ 0, & \text{if } \|b_j\| > \xi\sqrt{d}\lambda_1 \end{cases} = v_j$$

Solving the above equation, we have

$$\hat{b}_j = \begin{cases} \frac{\xi}{\xi-1} S_2(v_j, \sqrt{d}\lambda_1), & \text{if } \|v_j\| \leq \xi\sqrt{d}\lambda_1 \\ v_j, & \text{if } \|v_j\| > \xi\sqrt{d}\lambda_1 \end{cases} \quad (8)$$

where $S_2(v, \lambda) = \left(1 - \frac{\lambda}{\|v\|}\right)_+ v$. Thus to solve equation (7), we first need an initial value for \hat{b}_j . When $q \ll n$, a sensible initial value is the least squares estimate (without penalization). Then calculations described in this paragraph are iteratively carried out until convergence.

The overall algorithm is iterative. Within each iteration, we need to iteratively conduct GCD. For both the inner and outer iterations, we use the ℓ_2 norm of the difference between two consecutive estimates smaller than 10^{-5} as the convergence criterion. Convergence is achieved with all our simulated and real data. In the proposed objective function, the least squares term is continuously differentiable and regular in the sense of Tseng (2001). The penalty term is separable. Thus, the GCD algorithm converges to a coordinate-wise minimum of the first term, which is also a stationary point (Tseng 2001). In the outer iteration, each iteration leads to a decrease of the objective function. As the overall objective function is bounded from below, the proposed algorithm converges.

The proposed penalty contains two tuning parameters λ_1, λ_2 and one regularization parameter ξ . Generally speaking, smaller values of ξ are better at retaining the unbiasedness of the MCP penalties for larger coefficients, but they also have the risk of creating objective functions with a nonconvexity problem that are difficult to optimize and yield solutions that are discontinuous with respect to λ_1, λ_2 . It is therefore advisable to choose a ξ value that is big enough to avoid this problem but not too big. In our numerical study, we consider ξ values including 1.8, 3, 6, and 10, as in published studies. As with other penalization approaches, larger λ_1, λ_2 lead to fewer identified markers. In our numerical study, we jointly search for optimal $(\lambda_1, \lambda_2, \xi)$ values using V fold cross validation with $V = 5$. As with fixed tunings, the proposed algorithm only involves simple calculations, the proposed approach is

computationally feasible. For example, the analysis of one simulated dataset with $n = 250$ (details described in the next section) takes less than ten minutes on a regular desktop PC.

4 Numerical Study

In numerical study presented in this section, we focus on SNP data, which is categorical. The proposed approach can also be used to analyze continuous genetic data, for example, gene expressions. Numerical study with gene expression data is presented in Appendix.

4.1 Simulation

We conduct simulation to better gauge performance of the proposed approach. The SNP values are generated using a two-step approach. We first generate a 1000-dimensional vector with a multivariate normal distribution. The marginal means are equal to zero and marginal variances equal to one. We consider two correlation structures. The first is the auto-regressive correlation where the j th and k th components have correlation coefficient $\rho^{|j-k|}$. We consider $\rho = 0.2, 0.5, 0.8$, corresponding to weak, moderate and strong correlations, respectively. The second is the banded correlation. Here two scenarios are considered. Under the first scenario, the j th and k th elements have correlation coefficient 0.33 if $|j - k| = 1$ and 0 otherwise. Under the second scenario, the j th and k th elements have correlation coefficient 0.6 if $|j - k| = 1$, 0.33 if $|j - k| = 2$, and 0 otherwise. For each component of the 1000-dimensional vector, we dichotomize at the 1st and 3rd quartiles and generate the 3-level SNP value. For each subject, we simulate three clinical/environmental risk factors, with one continuously and two categorically distributed. For any two clinical/environmental risk factors, the correlation coefficient is 0.05. Under the simulated setting, there are a total of 1,003 ($3+1,000$) possible main effects and 3,000 (3×1000) interactions. Among them, six main effects and six interactions are set as associated with prognosis, all with regression

coefficients equal to 0.5. We generate the log event time from the AFT model with intercept equal to zero and standard normally distributed random error. The log censoring time is independently generated from a normal distribution. The censoring rate is about 40%. We simulate a total of 150 or 250 observations.

For comparison, we consider the following alternatives. Alt.1 is a marginal analysis approach as described in the Introduction section. Here it serves as benchmark. Alt.2 is a penalization approach, where the objective function is $L(\alpha, b) + \sum_{j=1}^p \sum_{k=1}^{q+1} \rho(|b_{jk}|; \lambda_2, \xi)$. Here we impose penalty on each main effect and interaction. This approach can conduct selection but does not respect the hierarchy of “main effect, interaction”. Alt.3 is also a penalization approach, where the objective function is $L(\alpha, b) + \sum_{j=1}^p \rho(\|b_j\|; \sqrt{d}\lambda_1, \xi)$. This approach only evaluates whether a SNP is associated with prognosis at all, but does not discriminate whether the association comes from the main effect or interaction. We are aware that a large number of approaches can be used to analyze the simulated data. The aforementioned three approaches are chosen for comparison as Alt.1 has been adopted in a large number of studies and can serve as benchmark. Comparison with Alt.2 and Alt.3 can directly establish the merit of each penalty term in formulation (4).

Summary statistics based on 100 replicates are shown in Table 1. Under all simulation settings, the proposed approach is able to identify the majority or all of the true nonzero interactions, while having a small number of false positives. Performance of the proposed approach depends on the correlation structure, strength of correlation, and sample size, as has been observed with other penalization approaches. The proposed approach also has satisfactory performance with the main effects, although the numbers of false positives are slightly larger than those with the interactions. The marginal approach (Alt.1) identifies a very small number of true positives. The unsatisfactory performance can be explained by the fact that interactions and main effects important in the joint model are not neces-

sarily important in the marginal models, especially when there exist complex correlations among variables. Alt.2 has satisfactory performance identifying the important main effects. However, without respecting the hierarchical structure, it identifies fewer true nonzero interactions compared with the proposed approach. Alt.3 identifies a large number of false interactions. Such an observation is expected, as under this approach, if a SNP has nonzero interaction effects with at least one clinical/environmental risk factor, it is concluded to have interactions with all clinical/environmental risk factors.

4.2 Analysis of an NHL prognosis study

NHL is a heterogeneous group of lymphocytic disorders ranging in aggressiveness from very indolent cellular proliferation to highly aggressive and rapidly proliferative process. It is the fifth leading cause of cancer incidence and mortality in the US and remains poorly understood and largely incurable. A genetic association study was conducted, searching for risk factors associated with overall survival in NHL patients (Zhang et al. 2004, Zhang et al. 2005). The prognostic cohort consists 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 (diffuse large B-cell lymphoma), 117 had FL (follicular lymphoma), 57 had CLL/SLL (chronic lymphocytic leukemia/small lymphocytic lymphoma), 34 had MZBL (marginal zone B-cell lymphoma), 37 had T/NK-cell lymphoma, and 96 had other subtypes. 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. In our analysis, we first analyze the whole cohort. In addition, we also analyze DLBCL, the largest subtype. Other subtypes are not analyzed because of sample size consideration.

When genotyping, we took a candidate gene approach. 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. We process data as follows. We remove patients with more than 20% SNPs missing and then remove SNPs with more than 20% measurements missing. The genotyping data were missing for the following reasons: the amount of DNA was too low, samples failed to amplify, samples amplified but their genotype could not be determined due to ambiguous results, or the DNA quality was poor. We then impute missing SNP measurements. A total of 1,633 SNPs pass processing, representing 238 genes.

For the whole cohort, 346 patients pass processing. Among them, 159 died, with survival times ranging from 0.04 to 11.01 years (mean 4.23 years). For the 187 censored patients, the followup times range from 4.85 to 11.50 years (mean 9.00 years). 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).

The following demographic and clinical factors were measured: age (rescaled to mean zero and variance one), education (level 1=high school or less; level 2=some college; level 3=college graduate or more), tumor stage (level 1-4 and unknown), B-symptom presence (no; yes; unknown), and initial treatment (none; surgery; radiation; chemotherapy; other). They include all widely accepted prognostic factors (Zhang et al. 2011).

We apply the proposed approach and the three alternatives described in the last section. In addition, we also employ the MCP penalization approach to the main effects only (additive effects of environmental and genetic risk factors). The difference in marker identification is summarized in Table 2. For all subtypes combined, the proposed approach identifies 11 main effects and 59 interactions. For DLBCL, it identifies 5 main effects and 24 interactions.

Table 2 suggests that the proposed approach identifies significantly different markers from the alternatives. Detailed results for the proposed approach are presented in Tables 3 (all subtypes) and 4 (DLBCL). Results for the other approaches are available from the authors. In Tables 3 and 4, the estimated regression coefficients are small. This is because the log-transformed event times are not rescaled. The magnitude of each interaction term is in general smaller than that of the corresponding main effect. However the sum of interaction coefficients is not necessarily smaller, suggesting that the hierarchical restriction in Bien et al. (2012) may be too strong. With the proposed approach, the dummy variables for a categorical clinical/environmental risk factor (for example tumor stage) are not selected as a whole. We intentionally design the proposed approach this way, so that it may identify which levels differ from the baseline. The proposed approach can be easily modified so that the dummy variables corresponding to the same risk factor can be selected together.

Searching published literature suggests that the identified genes may have important implications. For all subtypes combined, the protein encoded by gene MBL2 belongs to the collectin family and is an important element in the innate immune system. Deficiencies of this gene have been associated with susceptibility to autoimmune and infectious diseases. Polymorphisms of MBL2 have been associated with NHL in Mullighan et al. (2002). Mutations of MBL2 are suggested to be associated with follicular lymphoma patients' survival in Martinez-Lopez et al. (2009). Gene C4BPA encodes a member of a superfamily of proteins composed predominantly of tandemly arrayed short consensus repeats of approximately 60 amino acids. It is found to be significantly over-expressed in cancer patients with non-metastatic solid tumors (Battistelli et al. 2005). Protein encoded by gene CCR4 belongs to the G-protein-coupled receptor family. It is a receptor for the CC chemokines, which play fundamental roles in the development, homeostasis, and function of the immune system, and have effects on cells of the central nervous system as well as on endothelial cells involved

in angiogenesis or angiostasis. MASP1 encodes a serine protease that functions as a component of the lectin pathway of complement activation. The complement pathway plays an essential role in the innate and adaptive immune response. MIF is a ubiquitously expressed pro-inflammatory mediator that has also been implicated in the process of oncogenic transformation and tumor progression. Deletion of the MIF gene in mice has been shown to have several major consequences for the proliferative and transforming properties of cells. MIF-deficient cells exhibit increased resistance to oncogenic transformation (Fingerle-Rowseon and Petrenko 2007). The protein encoded by gene NCF4 is a cytosolic regulatory component of the superoxide-producing phagocyte NADPH-oxidase, a multicomponent enzyme system important for host defence. This gene belongs to the oxidative stress pathway, which is associated with NHL risk (Kim et al. 2012). The protein encoded by gene NOS1 belongs to the family of nitric oxide synthases, which synthesize nitric oxide from L-arginine. Nitric oxide is a reactive free radical, which acts as a biologic mediator in several processes, including neurotransmission, and antimicrobial and antitumoral activities. Gene SELP plays an important role in the pathogenesis of inflammation, thrombosis, and the growth and metastasis of cancers. Proteins encoded by genes STAT4 are members of the STAT protein family, which play important roles in lymphoma prognosis. Gene VCAM1 is a member of the Ig superfamily and encodes a cell surface sialoglycoprotein expressed by cytokine-activated endothelium. This type I membrane protein mediates leukocyte-endothelial cell adhesion and signal transduction.

Among the genes identified for DLBCL, the protein encoded by gene CCND1 belongs to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance throughout the cell cycle. This protein has been shown to interact with tumor suppressor protein Rb and the expression of this gene is regulated positively by Rb. Mutations, amplification and overexpression of this gene, which alters cell cycle

progression, are observed frequently in a variety of tumors and may contribute to tumorigenesis. The protein encoded by gene C5 is the fifth component of complement, which plays an important role in inflammatory and cell killing processes. Mutations in this gene cause complement component 5 deficiency, a disease where patients show a propensity for severe recurrent infections. The protein encoded by gene CCR7 is a member of the G protein-coupled receptor family. This receptor was identified as a gene induced by the Epstein-Barr virus (EBV), and is thought to be a mediator of EBV effects on B lymphocytes. This receptor is expressed in various lymphoid tissues and activates B and T lymphocytes. It has been shown to control the migration of memory T cells to inflamed tissues, as well as stimulate dendritic cell maturation. The protein encoded by gene RAC2 is a GTPase which belongs to the RAS superfamily of small GTP-binding proteins. Members of this superfamily appear to regulate a diverse array of cellular events, including the control of cell growth, cytoskeletal reorganization, and the activation of protein kinases. SOCS4 is a negative feedback regulator of EGF signalling, and has significantly attenuated expression in tumour tissue (Kobayashi et al. 2012).

In the literature, there is a lack of consensus on the genetic risk factors for NHL prognosis (Zhang et al. 2010). The study on $G \times E$ interactions is even more limited. Thus it is difficult to objectively assess the identification performance. Here we evaluate the prediction performance and reproducibility of the identified main effects and interactions (Huang and Ma 2010). To evaluate prediction performance, we randomly sample $3/4$ of the subjects without replacement and construct the training set. The corresponding testing set consists of the remaining subjects. We apply the proposed approach and analyze the training set. The training set model is then used to make prediction for subjects in the testing set. We dichotomize the testing set risk scores $\hat{\alpha}'X + \hat{\beta}'W$ at the median, create two risk groups, and compute the logrank statistic which measures the survival difference between the two

groups. To avoid an extreme sampling, we repeat the above process 500 times and compute the average logrank statistics. For all subtypes combined and DLBCL, the logrank statistics are 11.7 (p-value 0.0006) and 61.4 (p-value <0.0001), respectively, suggesting that the identified markers have satisfactory prediction performance. To evaluate the reproducibility performance, for a main effect/interaction, we count c , the number of times it is identified in the 500 resamplings. The proportion $c/500$, which has been referred to as occurrence index (Huang and Ma 2010), gives a measure on the relative importance and stability of the corresponding effect. For all subtypes combined, the identified main effects (presented in Table 3) have mean occurrence index 0.377, and the identified interactions have mean occurrence index 0.298. In contrast, the main effects and interactions not identified have mean occurrence indexes 0.021 and 0.015, respectively. For DLBCL, the identified main effects and interactions (presented in Table 4) have mean occurrence indexes 0.355 and 0.351, respectively. In contrast, the main effects and interactions not identified have mean occurrence indexes 0.028 and 0.020, respectively. Such results suggest the relative stability of the proposed approach and identified markers.

5 Discussion

In high-throughput cancer studies, it is of interest to identify $G \times E$ interactions which may be independently associated with prognosis beyond the main effects. In this study, for cancer prognosis data under the AFT model, we propose using a penalization approach for identifying $G \times E$ interactions. The proposed approach has an intuitive interpretation and is computationally feasible. Simulation study shows that it can outperform alternatives by identifying more true positives and fewer false positives. In data analysis, it identifies markers different from alternatives. The identified markers have important implications and satisfactory prediction performance and reproducibility.

In this article, we focus on prognosis data under the AFT model. It can be seen that the proposed penalty is relatively “independent” of the model setting. Thus it may be applicable with other data and model setup. We choose the AFT model because of its low computational cost and reasonable empirical performance. As there is a lack of model diagnostics tools for high dimensional data, comparison with other survival models is not pursued. We evaluate performance of the proposed approach using simulations. In data analysis, the identified genes are found to have important implications. However, we are unable to show that the proposed approach identifies “more meaningful” markers, as research on $G \times E$ interactions in NHL is very limited. The satisfactory prediction performance and reproducibility may provide partial support to the validity of the proposed approach. In the literature, a large number of studies have been devoted to identifying interactions. In this study, we focus on showing the feasibility of a penalization approach for cancer prognosis data. A thorough comparison with the existing methods will be postponed to future research.

References

- Amato R, Pinelli M, D’Andrea D, Miele G, Nicodemi M, Raiconi G, Coccozza S. (2010) A novel approach to simulate gene-environment interactions in complex diseases. *BMC Bioinformatics*. 11: 8.
- Battistelli S, Vittoria A, Cappelli R, Stefanoni M, Roviello F. (2005) Protein S in cancer patients with non-metastatic solid tumours. *Eur J Surg Oncol*. 31(7): 798-802.
- Bien J, Taylor J, Tibshirani R (2012) A Lasso for hierarchical interactions. [arXiv:1205.5050v1](https://arxiv.org/abs/1205.5050v1).
- Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, Fisher RI, Braziel RM, Rimsza LM, Grogan TM, Miller TP, LeBlanc M, Greiner TC, Weisenburger DD, Lynch JC,

- Vose J, Armitage JO, Smeland EB, Kvaloy S, Holte H, Delabie J, Connors JM, Lansdorp PM, Ouyang Q, Lister TA, Davies AJ, Norton AJ, Muller-Hermelink HK, Ott G, Campo E, Montserrat E, Wilson WH, Jaffe ES, Simon R, Yang L, Powell J, Zhao H, Goldschmidt N, Chiorazzi M, Staudt LM. (2004) Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *The New England Journal of Medicine*. 351: 2159-2169.
- Fingerle-Rowson G, Petrenko O. (2007) MIF coordinates the cell cycle with DNA damage checkpoints. Lessons from knockout mouse models. *Cell Division*. 2:22.
- Friedman J, Hastie T, Tibshirani R (2010) A note on the group Lasso and a sparse group Lasso. arXiv:1001.0736.
- Huang J, Ma S (2010) Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*. 16: 176-195.
- Huang J, Wei F, Ma S (2012) Semiparametric regression pursuit. *Statistica Sinica*, In press.
- Hunter DJ (2005) Gene-environment interactions in human diseases. *Nature Review Genetics*. 6: 287-298.
- Kim C, Zheng T, Lan Q, Chen Y, Foss F, Chen X, Holford T, Leaderer B, Boyle P, Chanock SJ, Rothman N, Zhang Y. (2012) Genetic polymorphisms in oxidative stress pathway genes and modification of BMI and risk of non-Hodgkin lymphoma. *Cancer Epidemiol Biomarkers Prev*. 21(5): 866-868.
- Kobayashi D, Nomoto S, Kodera Y, Fujiwara M, Koike M, Nakayama G, Ohashi N, Nakao A. (2012) Suppressor of cytokine signaling 4 detected as a novel gastric cancer suppressor gene using double combination array analysis. *World J Surg*. 36(2): 362-372.

- Martinez-Lopez J, Rivero A, Rapado I, Montalban C, Paz-Carreira J, Canales M, Martinez R, Sanchez-Godoy P, Fernandez de Sevilla A, Penalver FJ, Gonzalez M, Prieto E, Salar A, Burgaleta C, Queizan JA, Penarrubia MJ, Monteagudo MD, Cabrera C, De la Serna J, Tomas JF. (2009) Influence of MBL-2 mutations in the infection risk of patients with follicular lymphoma treated with rituximab, fludarabine, and cyclophosphamide. *Leuk Lymphoma*. 50(8): 1283-1289.
- Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol*. 241(2): 252-261.
- Mullighan CG, Heatley S, Doherty K, Szabo F, Grigg A, Hughes TP, Schwarer AP, Szer J, Tait BD, Bik To L, Bardy PG. (2002) Mannose-binding lectin gene polymorphisms are associated with major infection following allogeneic hemopoietic stem cell transplantation. *Blood*. 99(10): 3524-3529.
- North KE, Martin LJ (2008) The importance of gene-environment interaction: implications for social scientists. *Sociological Methods Research*. 37: 164-200.
- Stute W (1996) Distributional convergence under random censorship when covariates are present. *Scandinavian Journal of Statistics*. 23: 461-471.
- Thomas D (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health*. 31: 21-36.
- Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*. 109: 475-494.

- Witten DM, Tibshirani R (2010) Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*. 19: 29-51.
- Wu X, Jin L, Xiong M (2009) Mutual information for testing gene-environment interaction. *PLoS ONE* 4(2): e4578.
- Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*. 38: 894-942.
- Zhang Y, Dai Y, Zheng T, Ma S (2011) Risk factors of Non-Hodgkin lymphoma. *Expert Opinion on Medical Diagnostics*. 5: 539-550.
- Zhang Y, Holford TR, Leaderer B, Boyle P, Zahm SH, Flynn S, Tallini G, Owens PH, Zheng T. (2004) Hair-coloring product use and risk of non-Hodgkin's lymphoma: a population-based case-control study in Connecticut. *Am J Epidemiol* 159: 148-154.
- Zhang Y, Lan Q, Rothman N, Zhu Y, Zahm SH, Wang SS, Holford TR, Leaderer B, Boyle P, Zhang B, Zou K, Chanock S, Zheng T. (2005) A putative exonic splicing polymorphism in the BCL6 gene and the risk of non-Hodgkin lymphoma. *J Natl Cancer Inst* 97: 1616-1618.

Table 1: Simulation with SNP data: mean (standard deviation) based on 100 replicates.
TP/FP: true/false positives.

	Correlation	n	Main Effects		Interactions	
			TP	FP	TP	FP
Alt.1	AR $\rho = 0.2$	150	0.36(0.61)	0.03(0.17)	0.28(0.53)	0.67(1.56)
	AR $\rho = 0.5$	150	1.96(1.49)	1.59(10.55)	0.23(0.49)	1.44(7.72)
	AR $\rho = 0.8$	150	5.46(0.74)	0.63(3.63)	0.66(0.78)	0.86(1.68)
	AR $\rho = 0.2$	250	1.00(1.12)	0.11(0.82)	0.87(0.91)	2.36(14.48)
	AR $\rho = 0.5$	250	4.05(1.25)	5.18(50.49)	0.60(0.82)	3.05(17.53)
	AR $\rho = 0.8$	250	5.98(0.14)	1.45(5.79)	1.63(1.17)	2.22(4.74)
	Banded 1	150	0.60(0.97)	0.25(1.31)	0.18(0.41)	0.47(1.10)
	Banded 2	150	2.46(1.35)	0.03(0.30)	0.16(0.39)	0.66(1.37)
	Banded 1	250	1.75(1.40)	0.05(0.26)	0.70(0.80)	1.35(5.73)
	Banded 2	250	4.34(1.13)	0.09(0.55)	0.49(0.73)	1.05(3.11)
Alt.2	AR $\rho = 0.2$	150	5.74(0.54)	17.30(7.88)	3.61(1.29)	4.81(4.29)
	AR $\rho = 0.5$	150	5.53(0.66)	14.68(6.67)	3.72(1.39)	3.62(3.21)
	AR $\rho = 0.8$	150	4.72(0.71)	10.46(4.97)	3.32(1.11)	3.77(3.05)
	AR $\rho = 0.2$	250	5.98(0.14)	12.82(5.25)	5.27(0.78)	2.01(1.92)
	AR $\rho = 0.5$	250	5.90(0.30)	14.00(5.63)	5.30(0.73)	2.48(2.08)
	AR $\rho = 0.8$	250	5.29(0.52)	9.97(4.59)	4.69(0.66)	2.58(2.12)
	Banded 1	150	5.77(0.51)	15.97(6.14)	3.55(1.55)	3.91(3.37)
	Banded 2	150	5.14(0.85)	15.07(6.07)	3.11(1.57)	4.95(3.98)
	Banded 1	250	5.96(0.20)	13.88(6.01)	5.41(0.75)	2.26(1.82)
	Banded 2	250	5.70(0.48)	13.49(5.49)	4.91(0.85)	2.64(1.89)
Alt.3	AR $\rho = 0.2$	150	6.00(0.00)	12.12(4.76)	6.00(0.00)	48.36(14.28)
	AR $\rho = 0.5$	150	5.96(0.20)	13.34(5.29)	5.96(0.20)	51.94(15.85)
	AR $\rho = 0.8$	150	5.30(0.83)	12.62(5.47)	5.30(0.83)	48.46(17.00)
	AR $\rho = 0.2$	250	6.00(0.00)	8.25(6.86)	6.00(0.00)	36.75(20.58)
	AR $\rho = 0.5$	250	6.00(0.00)	8.66(5.60)	6.00(0.00)	37.98(16.80)
	AR $\rho = 0.8$	250	5.97(0.17)	14.16(6.73)	5.97(0.17)	54.42(20.30)
	Banded 1	150	5.98(0.14)	12.71(4.72)	5.98(0.14)	50.09(14.11)
	Banded 2	150	5.81(0.49)	14.96(5.20)	5.81(0.49)	56.50(15.79)
	Banded 1	250	6.00(0.00)	8.72(6.33)	6.00(0.00)	38.16(18.98)
	Banded 2	250	5.99(0.10)	12.24(7.12)	5.99(0.10)	48.70(21.39)
Proposed	AR $\rho = 0.2$	150	6.00(0.00)	19.34(6.81)	5.75(0.50)	4.40(3.01)
	AR $\rho = 0.5$	150	5.98(0.14)	19.75(6.92)	5.64(0.63)	4.64(3.35)
	AR $\rho = 0.8$	150	5.58(0.62)	20.10(6.10)	4.97(1.09)	6.10(3.53)
	AR $\rho = 0.2$	250	6.00(0.00)	11.23(8.89)	5.99(0.10)	2.36(3.48)
	AR $\rho = 0.5$	250	6.00(0.00)	10.88(7.51)	5.97(0.17)	2.12(2.50)
	AR $\rho = 0.8$	250	5.98(0.14)	14.53(8.58)	5.77(0.45)	3.28(3.57)
	Banded 1	150	5.99(0.10)	19.32(6.34)	5.83(0.43)	3.85(2.65)
	Banded 2	150	5.91(0.29)	21.92(6.72)	5.62(0.62)	5.19(3.27)
	Banded 1	250	6.00(0.00)	9.69(6.68)	5.95(0.22)	1.90(2.01)
	Banded 2	250	6.00(0.00)	11.85(8.51)	5.95(0.22)	2.33(3.17)

Table 2: Analysis of NHL data: main effects and interactions identified by different approaches and overlaps. In each cell, “a/b”=number of identified main effects/number of identified interactions. MCP: MCP penalization applied to the main effects only.

	MCP	Alt.1	Alt.2	Alt.3	Proposed
All subtypes combined					
MCP	23/0	2/0	3/0	1/0	8/0
Alt.1		3/0	1/0	0/0	2/0
Alt.2			9/36	0/3	1/5
Alt.3				4/52	3/26
Proposed					11/59
DLBCL					
MCP	13/0	0/0	2/0	1/0	1/0
Alt.1		0/1	0/0	0/0	0/0
Alt.2			2/9	1/1	0/0
Alt.3				4/52	1/9
Proposed					5/24

Table 3: Analysis of NHL overall survival, all subtypes combined: identified main effects and interactions.

SNP	Main	Age	Education		Interaction			
			Level 2	Level 3	Level 2	Level 3	Level 4	Unknown
MBL2_03	-0.002355	-0.000088			0.000059			-0.000294
ALOXE3_03	0.000147				-0.000031			0.000076
C4BPA_04	-0.003286	-0.000190						-0.000266
CCR4_01	0.000516	0.000184	-0.000023		-0.000486	-0.000051		0.000460
MASP1_69	-0.001931				0.000401			-0.000495
MIF_16	-0.001901	-0.000299		0.000374	-0.000784			-0.000809
NCF4_35	-0.000023				-0.000003			-0.000001
NOS1_18	0.000053		0.000040					0.000132
SELP_26	-0.001533							-0.001072
STAT4_33	-0.003080			0.000650				-0.000180
VCAM1_02	-0.000021	-0.000001				0.000001		-0.000010

SNP	B-Symptom			Initial Treatment		
	Yes	Unknown	Surgery	Radiation	Chemo	other
MBL2_03		0.001570				0.000749
ALOXE3_03						
C4BPA_04	0.000276	-0.000158	0.000144			
CCR4_01		0.000060		-0.000019	0.000458	
MASP1_69						
MIF_16	-0.000031		0.000808	0.000270	-0.000024	0.000115
NCF4_35				0.000005		0.000002
NOS1_18		-0.000013		-0.000006	0.000015	
SELP_26	-0.000184	0.000146	0.000078			
STAT4_33		0.000265	0.000655	0.000524		0.000260
VCAM1_02	0.000009	-0.000019	0.000004	0.000010	-0.000009	0.000005

Table 4: Analysis of NHL overall survival, DLBCL: identified main effects and interactions.

SNP	Main	Age	Education		Interaction			Unknown
			Level 2	Level 3	Level 2	Tumor Stage		
			Level 2	Level 3	Level 2	Level 3	Level 4	
CCND1_01	-0.001445	0.001385					-0.000800	
C5_15	-0.016346		0.010346	0.017137	0.002060			
CCR7_03	0.012630	0.008515	0.011905	-0.005183	-0.007506		0.010540	
RAC2_20	-0.000327			-0.000081				
SOCS4_01	0.000324	-0.000011					0.000113	

SNP	B-Symptom			Initial Treatment		
	Yes	Unknown	Surgery	Radiation	Chemo	other
CCND1_01				0.001084		
C5_15	0.009845					
CCR7_03		-0.009726		0.003497	0.007904	-0.000765
RAC2_20	0.000187	-0.000076		0.000027	-0.000052	
SOCS4_01			-0.000032			

Appendix

Here we conduct numerical study and evaluate performance of the proposed approach with continuously distributed genetic markers, in particular gene expression data.

Simulation

The simulation settings here are similar to those described for SNP data. The difference is that the dichotomization step is skipped. Thus, the simulated markers have multivariate normal distributions. Summary statistics based on 100 replicates are shown in Table 5. The observe patterns are similar to those in Table 1. In particular, the marginal approach fails to identify a satisfactory number of true positives. The penalization approaches are able to identify the majority or all of the true positives. The proposed approach outperforms Alt.2 and Alt.3 by identifying the smallest number of false positives.

Analysis of a follicular lymphoma prognosis study

Follicular lymphoma is the second most common form of NHL, accounting for about 22 percent of all cases. A study was conducted to determine whether the survival risks of patients with follicular lymphoma can be predicted by the gene expression profiles of the tumors and standard clinical risk factors at diagnosis (Dave et al. 2004).

Fresh-frozen tumor-biopsy specimens and clinical data from 191 untreated patients who received a diagnosis of follicular lymphoma between 1974 and 2001 were obtained. The median age at diagnosis was 51 years (range: 23 to 81), and the median follow up time was 6.6 years (range: less than 1.0 to 28.2). The median follow up time among patients alive at last follow up was 8.1 years. Eight records with missing survival information are excluded from the analysis. Clinical covariates measured include extra nodal site, age, normalized LDH, performance status, stage and IPI. We remove subjects with missing clinical covariate

measurements. 156 subjects are included in the analysis. Affymetrix U133A and U133B microarray genechips were used to measure gene expression levels. A log₂ transformation was first applied to the Affymetrix measurements. We select the 1,000 genes with the highest variations for downstream analysis. All gene expressions are normalized to have mean zero and variance one. In principle, there is no limit on the number of genes that can be analyzed using the proposed approach. We conduct prescreening to remove genes highly unlikely to be important to reduce computational cost and improve stability.

Summary of analysis results using different approaches is provided in Table 6. The proposed approach identifies 17 main effects and 29 interactions, which differ significantly from those identified using the alternative approaches. More detailed analysis results are provided in Table 7. The observed patterns are similar to those in Tables 3 and 4. We also evaluate the prediction performance and reproducibility using the approaches described for the NHL data. The prediction logrank statistic is calculated as 76.3 (p-value <0.0001), indicating satisfactory prediction performance. The identified main effects (presented in Table 7) have mean occurrence index 0.363, and the identified interactions have mean occurrence index 0.273. In contrast, the main effects and interactions not identified have mean occurrence indexes 0.021 and 0.014, respectively. We conclude reasonable reproducibility of the proposed approach and identified markers.

Table 5: Simulation with gene expression data: mean (standard deviation) based on 100 replicates. TP/FP: true/false positives.

	Correlation	n	Main Effects		Interactions	
			TP	FP	TP	FP
Alt.1	AR $\rho = 0.2$	150	0.24(0.55)	0.04(0.20)	0.14(0.35)	1.50(7.85)
	AR $\rho = 0.5$	150	2.27(1.55)	0.05(0.26)	0.17(0.40)	0.81(2.01)
	AR $\rho = 0.8$	150	5.77(0.49)	0.53(0.85)	1.39(1.09)	2.33(2.24)
	AR $\rho = 0.2$	250	1.20(1.30)	1.73(12.07)	0.59(0.74)	1.17(3.04)
	AR $\rho = 0.5$	250	4.36(1.11)	0.42(2.08)	0.77(0.87)	1.53(4.58)
	AR $\rho = 0.8$	250	5.99(0.10)	1.84(3.19)	2.74(1.20)	4.36(2.54)
	Banded 1	150	0.46(0.76)	0.40(3.32)	0.23(0.47)	1.21(2.90)
	Banded 2	150	2.82(1.34)	0.10(0.46)	0.06(0.24)	0.94(1.97)
	Banded 1	250	1.87(1.50)	0.11(0.49)	0.51(0.69)	3.24(20.38)
	Banded 2	250	4.42(1.04)	1.56(12.08)	0.43(0.71)	1.64(4.42)
Alt.2	AR $\rho = 0.2$	150	5.88(0.36)	8.37(3.53)	5.62(0.56)	16.16(4.60)
	AR $\rho = 0.5$	150	5.68(0.60)	7.69(3.25)	5.46(0.72)	15.10(5.10)
	AR $\rho = 0.8$	150	4.12(0.89)	4.99(2.95)	4.26(0.97)	12.60(5.15)
	AR $\rho = 0.2$	250	6.00(0.00)	6.87(3.29)	5.98(0.14)	16.47(6.03)
	AR $\rho = 0.5$	250	5.96(0.20)	6.54(3.67)	5.93(0.33)	16.29(6.72)
	AR $\rho = 0.8$	250	5.21(0.62)	6.35(3.15)	5.14(0.74)	16.75(6.85)
	Banded 1	150	5.82(0.44)	8.24(3.12)	5.48(0.78)	16.84(4.47)
	Banded 2	150	4.82(0.96)	6.80(3.25)	4.63(1.10)	15.16(5.39)
	Banded 1	250	6.00(0.00)	6.72(3.92)	5.97(0.17)	16.04(6.34)
	Banded 2	250	5.63(0.65)	7.64(3.74)	5.45(0.86)	19.57(7.38)
Alt.3	AR $\rho = 0.2$	150	6.00(0.00)	14.37(6.08)	6.00(0.00)	55.11(18.23)
	AR $\rho = 0.5$	150	6.00(0.00)	17.15(5.51)	6.00(0.00)	63.45(16.53)
	AR $\rho = 0.8$	150	5.32(0.82)	19.87(3.28)	5.32(0.82)	70.25(10.31)
	AR $\rho = 0.2$	250	6.00(0.00)	10.98(7.67)	6.00(0.00)	44.94(23.00)
	AR $\rho = 0.5$	250	6.00(0.00)	12.64(7.69)	6.00(0.00)	49.92(23.07)
	AR $\rho = 0.8$	250	5.91(0.29)	28.54(7.32)	5.91(0.29)	97.44(21.89)
	Banded 1	150	6.00(0.00)	15.55(4.92)	6.00(0.00)	58.65(14.75)
	Banded 2	150	5.73(0.62)	20.11(3.57)	5.73(0.62)	71.79(10.82)
	Banded 1	250	6.00(0.00)	11.31(7.40)	6.00(0.00)	45.93(22.19)
	Banded 2	250	6.00(0.00)	21.71(9.38)	6.00(0.00)	77.13(28.14)
Proposed	AR $\rho = 0.2$	150	6.00(0.00)	9.37(5.42)	6.00(0.00)	9.61(4.86)
	AR $\rho = 0.5$	150	5.98(0.20)	11.82(6.42)	5.97(0.22)	11.38(6.19)
	AR $\rho = 0.8$	150	5.32(0.84)	18.15(5.74)	5.10(0.96)	19.53(7.10)
	AR $\rho = 0.2$	250	6.00(0.00)	6.00(8.05)	6.00(0.00)	7.24(8.17)
	AR $\rho = 0.5$	250	6.00(0.00)	6.06(7.19)	6.00(0.00)	6.61(7.54)
	AR $\rho = 0.8$	250	5.94(0.28)	12.88(8.72)	5.87(0.37)	13.68(11.24)
	Banded 1	150	6.00(0.00)	11.53(6.40)	6.00(0.00)	10.79(6.13)
	Banded 2	150	5.76(0.61)	20.17(6.36)	5.58(0.81)	19.47(6.84)
	Banded 1	250	6.00(0.00)	5.29(6.76)	6.00(0.00)	6.18(7.46)
	Banded 2	250	6.00(0.00)	8.54(8.19)	6.00(0.00)	7.90(8.45)

Table 6: Analysis of follicular lymphoma data: main effects and interactions identified by different approaches and overlaps. In each cell, “a/b”=number of identified main effects/number of identified interactions. MCP: MCP penalization applied to the main effects only.

	MCP	Alt.1	Alt.2	Alt.3	Proposed
MCP	22/0	0/0	12/0	2/0	10/0
Alt.1		0/0	0/0	0/0	0/0
Alt.2			14/12	2/1	8/2
Alt.3				16/96	6/14
Proposed					17/29

Table 7: Analysis of follicular lymphoma data: identified main effects and interactions.

Feature ID	Gene	Main effect	Nodal	Age	Interaction			Stage	IPI
					LDH	Pstat			
224114_at	MAP2K5	0.013755						-0.003937	
227860_at	PCED1A	0.000847	-0.000018						
231385_at	DPPA3	0.002453		-0.001007			0.003739		
231495_at	VTI1A	0.000511				-0.000053			
237007_at	KCNB2	-0.008013						-0.000071	
238584_at	FLJ22527	0.000709				-0.000017	0.000487	-0.000160	
242334_at	NALP4	0.010591		-0.001079				-0.000777	
224101_x-a	MRS2L	-0.001058				0.000528			
230894_s-a	MSI2	0.014455	-0.005527			-0.006882			
207245_at	UGT2B17	0.000508	-0.000100						
211736_at	SP2	0.013484		-0.009329				-0.005664	
215962_at	SNTG1	0.001579					0.000017		
216686_at	FLJ40330	-0.005010	-0.001806	0.000730					
217469_at	IGHE	0.002981	-0.000548	-0.000366					
204463_s-a	EDNRA	0.012517						-0.005470	
208168_s-a	CHIT1	-0.001142	0.000001					-0.000001	
209631_s-a	GPR37	0.006800	-0.003217	-0.000107				0.000593	