Genomic data analysis using the L_1 and L_2 penalized regression method in cancer clinical researches for establishing the precision medicine

(がん臨床研究における罰則付き回帰分析
を用いた遺伝子データ解析法に関する研究)

Shuhei Kaneko

Tokyo University of Science

March 2017

TABLE OF CONTENTS

Preface

The precision medicine is defined as ”an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person”.[1] Genomic technologies are rapidly developed, and allows us to investigate the condition in human cell by quantifying the level of gene expression, measuring the mutation for a large number of gene, and so on. These genomic data are frequently used for several types of cancer clinical researches such as the prognostic model developments for cancer patient, and developments for novel and evolutional drug for establishing the precision medicine.

In general, the genomic data contains a lot of genes while a number of patients is limited. In this setting, the traditional statistical approach cannot work well because it generally requires much more patients than genes. To address this difficulty, researchers have been emphasizing the penalized regression methods. Among them, the L_1 (with or without L_2) penalized regression, which select important genes for prediction and simultaneously estimate the regression coefficients, is a typical and frequently used penalized regression method in cancer clinical researches. This method shrinks all regression coefficients toward zero, and automatically sets many of them to exactly zero, depending on the amount of regularization employed. The penalized regression approach is useful for almost all kinds of cancer clinical researches. However, several issues exist when we utilize the penalized regression model depending on the types of researches and types of datasets used.

In this study, we clarify two focal types of clinical researches and the corresponding issues about utilizing the L_1 and L_2 penalized regression method as follows:

Issue 1

In the past decade, researchers in oncology have sought to develop survival prediction models using high-dimensional gene expression data. The least absolute shrinkage and selection operator (lasso) has been widely used to select genes that truly correlated with a patient’s survival. The lasso selects genes for prediction by shrinking a large number of coefficients of the candidate genes towards zero based on a tuning parameter that is often determined by a cross validation (CV).

The CV method determines the value of the tuning parameter by considering the trade-off between the number of true positives and false positives in selected genes, and so the possibility of (i) containing false positives and (ii) identifying false negatives cannot be eliminated.

(Issue 1-1) We propose a method for estimating the false positive rate (FPR) for lasso estimates in a high-dimensional Cox model. We performed a simulation study to examine the precision of the FPR estimate by the proposed method. We applied the proposed method to real data and illustrated the identification of false positive genes.

(Issue 1-2) The CV can pass over (or fail to identify) true positive genes (i.e. it identifies false negatives) in certain instances, because the lasso tends to favor the development of a simple prediction model. We attempt to monitor the identification of false negatives by developing a method for estimating the number of true positive (TP) genes for a series of values of a tuning parameter that assumes a mixture distribution for the lasso estimates. Using our developed method, we performed a simulation study to examine its precision in estimating the number of TP genes. Additionally, we applied our method to a real gene expression dataset and found that it was able to identify genes correlated with survival that a CV method was unable to detect.

Issue 2

The precision medicine is defined as ”an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person”.[1] Cancer is a genomic disease, and so molecularly targeted agents (MTAs) for cancer recently developed are thought to be that the dose-efficacy and dose-toxicity relationships differ depending on the gene mutation pattern. The individualized optimal dose (IOD), which is defined as the maximal efficacious dose which can be administered with clinically acceptable safety profile varying depending on the gene mutation patterns, should be determined for MTAs for establishing the precision medicine. In addition, the genes which determine the IOD for MTAs should be identified in early phase of developments. We propose a novel dose-finding approach to identify the IOD for MTAs in phase I trials in oncology. An IOD determination and gene selection are simultaneously performed based on the L_1 and L_2 penalized regression. Many dose-finding approaches for MTAs in the available literature account for the non-monotonic patterns for dose-efficacy and dose-toxicity relationships as well as correlations between efficacy and safety outcomes, and we consider them by the penalized regression based on the multinomial distribution. The dose-finding algorithm is based on the predictive values which are calculated by the estimated penalized regression model. We compare the operating characteristics between the proposed and existing methods by simulation studies under various scenarios.

This doctoral dissertation consists of six chapters. Chapter 1 provides a background of cancer prognostic studies as well as the molecularly targeted agents related to the issue 1 and 2 in cancer clinical researches, and an outline of the penalized regression model for the Cox proportional model and the generalized regression model. The main topic of Chapter 2 is to propose a mixture distribution for the lasso estimates to estimate the FPR in a high-dimensional Cox model. This chapter is based on Kaneko et al (2012)[2]. In Chapter 3, we propose to monitor the identification of false negatives by developing a method for estimating the number of true positive (TP) genes for a series of values of a tuning parameter by utilizing the proposed mixture distribution in Chapter 2. This chapter is based on Kaneko et al (2014)[3]. In Chapter 4, the main concern is to propose an individualized dose-finding approach using the penalized regression model with simultaneous gene selection in early phase developments of molecularly targeted agents. This chapter is entirely based on Kaneko et al (In preparation)[4]. Finally, Chapter 5 discusses the issues relating this study and Chapter 6 presents the conclusions of this study.

Chapter 1

Introduction

1.1 General background of this study

Cancer is a disease of which development is that cancer cells abnormally increase and invade surrounding tissues and organs, and subsequently, through the blood and lymphatic vessel, spread to distant organs.[5] Cancer is caused by several factors; tobacco, infectious organisms, and an unhealthy diet as factors outside a body, and genetic mutations, hormones, and immune conditions as factors inside a body.[5] Without proper treatments for controlling the cancer, it may result in severe condition including death. Several types of treatments exist; surgery to remove cancer, radiation, and therapies based on drugs which include chemotherapy, hormone therapy, and targeted therapy.[5] It is known that a cancer is a disease of genome, and many of the molecular lesions has its own genomic feature, some of which are in common among multiple types.[5]

In Japan, cancer has been a leading cause of death since 1981.[6] In 2014, 368,103 patients were dead by cancer, and five leading sites in mortality are; lung, stomach, colon/rectum, pancreas, and liver for both sexes; lung, stomach, colon/rectum, liver, and pancreas for males; and colon/rectum, lung, stomach, pancreas, and breast for females.[6]

There have been a progress to fight against cancer in the past decades. According to the American cancer society[5], in America death rate by any cancer had increased by late of the 20th century because the tobacco was taken by a lot of people (in 1991 at 215 cancer deaths per 100,000 persons at peak time); however, from the peak time to 2012, the rate decreased by 23% (the decrease of more than 1.7 million cancer deaths). Death rates are declining for all four of the most common cancer types; lung, colorectal, breast, and prostate.[5] In Japan, the similar tendency has been shown in the same period. The Cancer statistics in Japan 2015[6] reported that for males the age-adjusted rates of cancer mortality (all ages) increased by around 1990 and reached a peak, and afterwards has been decreasing since late of the 20th century while for females it has been gradually decreasing since around 1970. The age-adjusted rates of cancer mortality of both sexes gradually decreased since around 1960 to around 1990 and has been clearly decreasing since late of the 20th century.[6]

Obviously, developments of the novel treatments for cancer in this decade have contributed to this improvement in patients with cancers. Treatments are evaluated in human in a series of steps of clinical trials; phase I, II, and III, before they are commercially available. Trials in early phases are exploratory stage to evaluate the efficacy and the safety for candidate drugs and especially in cancer clinical trials to determine the optimal dose for subsequent phases, with relatively small number of patients (or sometimes healthy volunteers in clinical trials other than treatments for cancers) while the trials in later phase are to confirm the efficacy and safety compared with the placebo or the active comparator with relatively large number of patients. The target patient population of the drug in the trials is often based on the whole population of patients with specific disease, for example an angiotensin II receptor blocker for patients with hypertension. This approach was taken over to developments of treatments for cancers. However, it is known that successes rate of development based on this approach have been low.[7] One report showed that, since 1993 to 2004, only 13% of cancer drugs of which developments were initiated in phase I trials were finally approved by the US Food and drug administration (FDA) as regulated products.[8] Another report showed that, since 2003 to 2011, only 10.5% of newly developed agents were finally approved by the FDA and most of them failed in phase II trials.[9] This low rate of the success drove the paradigm shift from the traditional approach to developments for the treatments targeted to more specific population, for example treatments targeted towards specific molecules which involves in cancer progression or metastasis.[7]

The precision medicine is defined as ”an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person”.[1] The benefit of the precision medicine would be seen in oncology at first because, as written above, a cancer is a leading cause of death, the success rate in developments is low, cancer is a disease of genome, and also the breakthrough treatments are still required for patients in severe conditions such as patients with recurrent or metastatic cancers. The recent developments in oncology have been already along with the paradigm of the precision medicine. The trastuzumab was developed for the treatment of breast cancer and gastric cancer with over expression of HER2 gene. This drug was created based on the knowledge that HER2 is possibly related to proliferation of cancer cells, and subsequent clinical trials have been conducted in patients with over expression of HER2 genes.[10] There is another example. The Oncotype DX® Recurrence Score® is based on expression levels of 21 genes and approved by FDA as diagnosis for predicting the chemotherapy benefits and likelihood of the recurrence for patients with breast cancers.[11] These kinds of approaches will further emerge in near future in the paradigm of the precision medicine.

Recent developments of technologies in biology contributed to establishing the precision medicine. The array-based hybridization assay, such as the DNA microarray, and more advanced technology including the next generation sequencing allow us to acquire tens of thousands of genomic characteristics, such as the expression levels of large number of genes or to genotype for multiple regions of a genome, with high quality and much cheaper price than the past. These technologies facilitate to identify important molecular targets of treatments as biomarkers, some of which have been already used for diagnosis.

There are two types of biomarkers as Simon[12] reported. Prognostic markers are baseline measurements that provide whether the patient probably survive long-term or not with either untreated or a standard treatment, which are utilized to determine whether any systematic treatment or any therapy beyond the standard treatment is required or not. Predictive markers are baseline measurements that indicate whether the patient is likely (or unlikely) to benefit from a specific drug or regimen. The use of these biomarkers is potentially useful for establishing the precision medicine in oncology through developments of diagnosis (eg, identifying patients who had better prognosis or better treatment effect, adjusting dose based on the biomarkers for “right dose for right patient”) and developments of the novel treatment, such as molecularly targeted agents that inhibit the working of a specific molecule.

1.2 Statistical methods for establishing the precision medicine

The statistical challenges exist in analyses of genomic data. The genomic data consist of numbers of genes from relatively fewer patients. The number of genes included in the statistical model vary depending on researches, but may be tens of thousands of genomic characteristics at most. Ordinal regression model cannot be directly applied since the number of patients much exceeds the number of genes in most cases. Sometimes researchers could pick up limited numbers of candidate genes (e.g., 4 or 5) for analyses based on the prior knowledge. However, the predictive biomarkers are identified by using the models with an interaction between treatment and candidate biomarkers. So, if researchers want to explore the predictive biomarkers, they suffer from number of covariates of genes and interactions, and the statistical challenge still remains even though the genes have been already picked up.

Several researchers addressed this challenge. A simple approach such as the univariate regression method was proposed in the early era of genomic data analyses. Subsequently, more advanced and sophisticated approaches based on the dimension reduction or penalized regressions were proposed.[13–20] Bovelstad et al[21] systematically compared the operating characteristics between these methods which included univariate regression method, forward stepwise selection methods, principal components regression (refer to Hastie et al.[13]), supervised principal components regression[14,15], partial least squares regression[16], ridge regression which is L_2 penalized regression[17], and least absolute shrinkage and selection operator (lasso) which is L_1 penalized regression[18,19], based on large number of gene expression data from cancer patients as well as their survival times. They found that the prediction ability of univariate and forward stepwise selection was much worse than the one of the other advanced and sophisticated methods, and the ridge regression was the best method in terms of prediction. One drawback of the ridge regression approach is to include all genes (eg, tens of thousands) in the model. In many studies, however, to identify a small subset of the genes is one of the most important objective. The ridge regression does not have a property of gene selection. Genes identified by the statistical methods are subsequently investigated further in order to obtain a mechanistic understanding from biological point of view. With these purposes considered, they concluded the lasso is one of the most interesting methods in genomic data analyses in cancer clinical researches with acceptable level of decrease in prediction ability.[21] More recently, Zou et al[20] proposed the L_1 and L_2 penalized regression called as the elastic net which is the most promising extensions of the lasso, and has the same property of the lasso based on the L_1 penalty and a more favorable property for correlated covariates than the lasso. From these points, in this paper, we focus on the penalized regression methods of the lasso and the elastic net in the rest of the paper.

1.3 Cancer prognostic studies

Establishing prognoses of clinical outcomes on the basis of gene expression data is often performed in this decade.[22–25] In cancer clinical researches, not only the prediction of response to treatment but also the prediction of time to such events, e.g., overall survival (OS) and relapse-free survival (RFS) are investigated.[26] To precisely predict such outcomes, we need to identify the genes that are highly correlated with them and are called the outcome-predictive genes. This is difficult, however, because the number of genes p in the high-dimensional gene expression data exceeds the number of patients n. Several researchers have attempted to identify the outcome-predictive genes in the $n < p$ data settings by using traditional statistical methods, but the accuracy of the prediction based on the genes identified in this way is not very satisfactory. For example, van't Veer et al[23] and Van de Vijver et al[24] analyzed the gene expression profiles of 78 lymph node-negative breast cancer patients in order to establish gene signatures related to the risk of distant metastasis. Using a "three-step supervised classification method", they identified 70 genes that categorize patients into "good" and "bad" prognostic groups. Wang et al[25] also analyzed the gene expression profiles of 115 patients for the same purpose. They identified 76 genes by using the univariate regression of Cox's proportional hazard regression analysis[27], which evaluates the relationship between the level of expression and the distant-metastasis-free survival for each gene. Notably, both studies had only 3 genes in common. Furthermore, the predictive performance based on both gene signatures drastically decreased when applied to other data sets.[28] Thus, the problem lies in the difficulty of precise identification of the outcome-predictive genes in high-dimensional data.

Table 1.1: Cancer clinical researches for gene expression data

	van't Veer et al[23]	Wang et al[25]
Objective	To select the outcome-predictive genes and develop the prediction model	
Number of patients	78	115
Number of genes	5,000	17,819
Method	Correlation	Univariate Cox model
Number of selected genes	70	76
Number of selected genes in common for both studies	3	

1.3.1 The L_1 penalized regression

Researchers have been emphasizing the penalized regression methods. Among them, the least absolute shrinkage and selection operator (lasso), which selects the outcome-predictive genes and simultaneously estimates the regression coefficients in the Cox regression model, is a typical penalized regression method.[18,19] This method shrinks all regression coefficients toward zero, and automatically sets many of them to exactly zero, depending on the amount of regularization employed. This can be useful, in particular, in high-dimensional data, and the prediction performance for gene expression data have been widely studied by many researchers by using this method.[21,29,30] Several researchers showed that the lasso outperforms the simple variable selection methods such as the univariate Cox regression analysis,[21,31] with respect to the accuracy of prediction.

The lasso shrinks most of the coefficients towards zero exactly by adding L_1 norm to the Cox log partial likelihood, and the amount of shrinkage is dependent on the tuning parameter. The value of the tuning parameter is often determined by a cross-validation (CV), which maximizes the out-of-data prediction accuracy.[32]

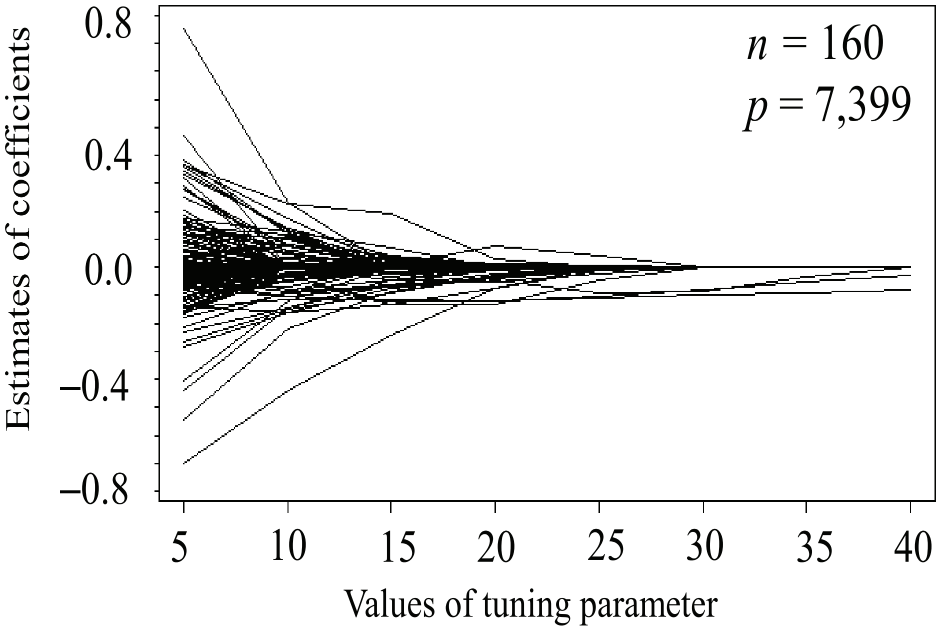

Figure 1.1: A result applying the lasso to the gene expression dataset published by Rosenwald et al[22]. n: a number of patients, p: a number of genes

Several researchers have investigated the operating characteristics of the lasso. Goeman[33] used the lasso to analyze a publicly available gene expression dataset, obtained from the articles of van't Veer et al.[23] and van de Vijver et al.[24] in which a 70-gene signature for prediction of metastasis-free survival in breast cancer patients had been established. This data included 295 patients with 4,919 genes that were prescreened from 24,885 genes based on the quality criteria in van't Veer et al.[23]. The lasso selected 16 genes with which to develop a prediction model of overall survival when using the tuning parameter that was determined using a CV. Goeman[33] also conducted ridge regression using all 4,919 genes to develop a model by adding L_2 norm to the Cox log partial likelihood. The prediction accuracy of the lasso and ridge regression were compared, and the ridge regression with 4,919 genes slightly outperformed the lasso with 16 genes. Goeman[33] suggested that the lasso potentially passes over genes that are correlated with survival in order to develop a simple prediction model. Bøvelstad et al.[21] reached the same conclusion in a review of the survival prediction methods available for analyzing breast cancer gene expression datasets. Table 1.2 summarizes a typical result of gene selection by the lasso.

Table 1.2: Typical results of gene selection by the lasso

True condition	The lasso	
	Select	No select
Genes that are NOT correlated with survival (none-outcome-predictive genes)	False positive (FP)	True negative (TN)
Genes that are truly correlated with survival (outcome-predictive genes)	True positive (TP)	False negative (FN)

The CV method determines the value of the tuning parameter by considering the trade-off between the number of true positives (TP) and false positives (FP), and so the possibility of identifying false negatives (FN) cannot be eliminated. To further investigate the operating characteristics of the lasso with CV regarding the number of TP and FP, we conducted simulation studies with assuming typical analyses of gene expression data (see Appendix A). The simulation study demonstrated that the lasso possibly fail to identify true set of the TP with including a lot of number of FP.

One solution for correctly identifying outcome-predictive genes is at first to monitor the number of TP in several values of the tuning parameter and determine its final value, and subsequently remove the FP. For issue 1, we proposed methods to achieve this objective.

1.3.2 Objective of this study (Issue 1)

(Issue 1-1) We proposed a method for estimating the false positive rate (FPR) for lasso estimates in a high-dimensional Cox model. We performed a simulation study to examine the precision of the FPR estimate by the proposed method. We applied the proposed method to real data and illustrated the identification of false positive genes.

(Issue 1-2) We developed a method for estimating the number of TP for a series of values of the tuning parameter. We assumed a mixture distribution for the lasso estimates developed in the Issue 1-1, and these could be used to estimate the number of TP and FP. It is possible to generate the solution path that includes the lasso estimates for a series of values of the tuning parameter using the methods developed by Goeman[33]. Here, we proposed an algorithm to sequentially fit the mixture distribution for this solution path, and we used a simulation study to test the precision of the algorithm when estimating the number of TP. We further demonstrated the proposed algorithm using a well-known diffuse large B-cell lymphoma (DLBCL) dataset comprising overall survival of 240 DLBCL patients and gene expression data of 7,399 genes.[22]

1.4 Clinical developments for novel molecularly targeted agents

1.4.1 Molecularly targeted agents and dose-finding

Developments of molecularly targeted agents (MTAs) for cancer which inhibit the working of a specific molecule (e.g., small molecules or monoclonal antibodies) has been initiated in an effort to move toward the precision medicine, which is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.[1] The phase I trials for MTAs have been conducted to find the optimal dose which is used for later phase of developments for overall population.[34–38,40] To establish the precision medicine, the individualized optimal dose (IOD), which is defined as the maximal efficacious dose which can be administered with clinically acceptable safety profile varying depending on the patients' characteristics, should be determined for MTAs in phase I trials. In addition, the patients' characteristics which determine the IOD for MTAs should be identified in early phase of developments for further researches.

1.4.2 Motivating example

One motivating example is a dose-finding trial for the IOD of a novel MTA in patients with locally advanced or metastatic solid tumors, which was reported in Guo et al.[39] Five genes (genes considered in the trial were the NTRK1 gene, NTRK2 gene, NTRK3 gene, ROS1 gene, and ALK gene) which had been considered that the benefit and risk of the treatment would differ depending whether the gene mutation occur or not. The mutation status of each of the five target genes was coded as positive or negative which was measured by using the next-generation sequencing[40]. The efficacy and toxicity outcome for each patient are evaluated on the graded scale based on the National Cancer Institute Common Terminology Criteria for Adverse Events and the Response Evaluation Criteria In Solid Tumors, respectively. Five doses are evaluated. For this trial, the most simple approach to model the dose-efficacy and dose-safety relationships is to include dose and genes as covariates for main terms and dose by genes as covariates for interaction term in the dose-toxicity and dose-efficacy models. However, the model cannot work well in many times because of small sample size in contrast to a lot of number of covariates caused by including genes in main terms and gene by treatment in interaction terms. Moreover, genes could be moderately or highly correlated.

1.4.3 The L_1 and L_2 penalized regression

Zou et al[20] proposed the L_1 and L_2 penalized regression, which is called as the elastic net, by adding L_2 norm to the penalty function of the lasso. The lasso sometimes select only one covariate from correlated covariates or all covariates with very small values of estimates. In addition, highly correlated covariates may yield the unstable regularization path of the lasso estimates. The form of the penalty based on both the L_1 and L_2 can address the problems of the lasso as reported in Friedman et al.[41] The L_2 penalized regression, which is called as the ridge regression[17], is known to shrink the coefficients of correlated covariates towards each other. From a Bayesian point of view, the ridge penalty shows the good property in case that many covariates exist in the model, and all have non-zero coefficients as drawn from a normal distribution with mean = 0 as prior distribution. The lasso estimates are considered as the posterior mode with a Laplace prior, which expects many coefficients to be shrunken to exactly zero, and a small subset of estimates to be nonzero (ie, the lasso has the property of the gene selection while the ridge regression does not have). The ratio of amounts of L_1 and L_2 penalties are determined by the mixing-control parameter in the elastic net. When the mixing-control parameter suggests to use both L_1 and L_2 penalties, the elastic net performs like the lasso, but removes any drawbacks caused by highly correlated covariates. This property is much useful in the setting of this issue 2 because the model utilized includes multiple mutations of moderately or highly correlated genes as well as both main and interaction terms in the model which may result in high correlation in the model. (However, note that the elastic net has more challenges in choosing the values of the two tuning parameters and the mixing-control parameter appropriately while only one tuning parameter exist in the lasso.)

1.4.4 Objective of this study (Issue 2)

We proposed a novel dose-finding approach to identify the IOD for MTAs in phase I trials in oncology. We utilize the L_1 and L_2 penalized regression, which is one of the most frequently used penalized regression in analyses of the high-dimensional data and so called as the elastic net[20], and the IOD determination and gene selection are simultaneously performed based on the elastic net. For each binary efficacy and toxicity outcome, we assume the logistic model with including dose and quadratic form of dose, and gene effect as covariates in main term, and dose by gene effect as covariates in interaction term. The quadratic form of dose in the model is incorporated to account for the non-monotonic patterns for dose-efficacy and dose-toxicity relationships as Cai et al.[42] The estimation of the coefficients for efficacy and toxicity outcome are performed separately and based on the frequentist approach by maximizing the penalized log-likelihood. The dose-finding algorithm is based on the predictive values which are calculated by the estimated penalized regression model. Many dose-finding approaches for MTAs in the available literature take account into correlations between efficacy and toxicity outcomes, and we also consider them based on the multinomial distribution for bivariate joint binary probability of efficacy and toxicity outcome as Sato et al[36]. We implement the simulation studies, and compare the operating characteristics between the proposed method and the method of Wages and Tait[38] with executing the elastic net only once at the end of the trial.

Chapter 2

Gene selection using a high-dimensional regression model with gene expression data in cancer prognostic studies

2.1 Introduction

In this chapter, we developed a method for estimating the proportion of FP genes, i.e., false positive rate (FPR), among the total identified genes. Specifically, the FPR is calculated using a mixture distribution based on the coefficients estimated by the lasso. We formulate the mixture distribution by considering the features of the lasso. By identifying the FP genes using the proposed method and excluding them from the Cox model, we are able to improve the prediction accuracy of the model. The accuracy of the FPR estimated by the proposed method is examined by simulation studies. We present the illustration of the proposed method using a well-known data set containing gene expressions from patients with diffuse large B-cell lymphoma (DLBCL) along with their survival time.[22]

2.2 Method

Cox proportional hazard model

Consider a sample of size n from which the relationship between the timing of an event and gene expression levels $x_1, \ldots, x_p$ of p genes need to be estimated. Due to censoring, for $i = 1, \ldots, n$, the ith datum in the patient is denoted by $(t_i, \delta_i, x_{i1}, \ldots, x_{ip})$, where δ_i is the censor indicator and t_i is the event time if $\delta_i = 1$ or censored time if $\delta_i = 0$, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the vector of the gene expression levels of p genes for the ith patient. The Cox proportional hazard model is the most popular method to evaluate the relationship between gene expression and survival outcomes.[27] The hazard function of an event at time t for a patient i with the gene expression levels $\mathbf{x}_i$ is given by

$$h(t|\mathbf{x}) = h_0(t) \exp\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a parameter vector and $h_0(t)$ is the baseline hazard, which is the hazard for the respective individual when all variable values are equal to zero. In the general setting with $n > p$, the coefficients are estimated by maximizing Cox's log partial likelihood as follows:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \log\left\{\sum_{r \in R(t_i)} \exp\left(\mathbf{x}_r^{\mathrm{T}} \boldsymbol{\beta}\right)\right\}\right] \tag{2.2}$$

where $R(t_i)$ is the risk set that contains the patients whose survival time or censored time is at least t_i.

The lasso

Tibshirani[18,19] introduced a novel parameter estimating method that simultaneously executes parameter estimation and variable selection by adding the L_1 norm to log likelihood function. The penalized log-likelihood function l_p of the lasso in the Cox's proportional hazard model is as follows:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.3}$$

where λ is the tuning parameter, which determines the amount of shrinkage, and $l(\boldsymbol{\beta})$ is the Cox's log partial likelihood. The parameters are estimated by maximizing Equation (2.3). In this study, the parameters were estimated using the efficient gradient ascent algorithm.[33]

When performing the lasso, we need to determine the value of λ, which affects the lasso estimates. As the value of λ increases, the number of the selected genes monotonically decreases. The optimal value is often determined by the cross-validation log partial likelihood[32]. The K-fold cross-validated log partial likelihood is given by

$$CV(\lambda) = \sum_{k=1}^{K} \left\{ l\left(\hat{\boldsymbol{\beta}}_{(-k)}\right) - l_{(-k)}\left(\hat{\boldsymbol{\beta}}_{(-k)}\right)\right\} \tag{2.4}$$

where $l_{(-k)}\left(\hat{\boldsymbol{\beta}}\right)$ is the log partial likelihood when the k-th fold is left out, and $\hat{\boldsymbol{\beta}}_{(-k)}$ is the estimate of $\boldsymbol{\beta}$ obtained by the lasso when the k-th fold is left out. The optimal tuning parameter λ is obtained by maximizing $CV(\lambda)$. The number of folds to execute the above-mentioned cross validation is often set to 5 (or 10), considering the computational feasibility.

Estimation of false positive rate (FPR)

In this section, we propose the method to estimate the FPR for a fixed value of λ determined by the cross validation by assuming a mixture distribution for the lasso estimates. The mixture distribution is developed on the basis of the following 2 features of the lasso: (i) the lasso selects at most n variables, because of the nature of the convex optimization problem when $n < p$,[20,43] and (ii) in the Bayesian framework, the lasso estimate is derived as the posterior mode under independent Laplace prior distribution as follows:

$$f_L\left(\beta_j; 0, \frac{1}{\lambda}\right) = \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right) \tag{2.5}$$

where $f_L\left(y; a, b\right) = (2b)^{-1} \exp\left(-|y - a|/b\right)$ is the probability density function of Laplace distribution with location parameter a and scale parameter b.[18] On the basis of these features of the lasso, the mixture distribution is assumed to $\hat{\beta}_j$ for the fixed value of λ as follows:

$$\begin{aligned} f\left(\hat{\beta}_j; \pi_0, \pi_c, \tau, \mu_c, \sigma_c\right) = {} & \tfrac{n}{p}\left\{\pi_0 f_L\left(\hat{\beta}_j; 0, \tau^{-1}\right) + \textstyle\sum_{c=1}^{C} \pi_c f_N\left(\hat{\beta}_j; \mu_c, \sigma_c^2\right)\right\} \\ & + \left(1 - \tfrac{n}{p}\right) f_L\left(\hat{\beta}_j; 0, \epsilon\right) \end{aligned} \tag{2.6}$$

where π_0 and π_c are mixed proportions $\left(\pi_0 + \sum_{c=1}^{C} \pi_c = 1\right)$; $f_N\left(\cdot; \mu_c, \sigma_c^2\right)$ is the probability density function of the normal distribution with mean μ_c ($\neq 0$) and variance σ_c^2 in component c; C is the number of component, which is determined on the basis of any model evaluation criteria; and ϵ is the constant value, which is boundlessly close to 0, e.g., $\epsilon = 10^{-8}$. The unknown parameters, π_0, π_c, τ, μ_c, and σ_c, are estimated by maximizing the log-likelihood function of Equation (2.6).

The mixture distribution defined by Equation (2.6) is formulated on the basis of the following concepts. Since the lasso selects at most n genes in the $n < p$ setting, the coefficients for at least $p - n$

genes are shrunken toward exactly zero; therefore, Equation (2.6) consists of 2 terms, i.e., n/p term and $1 - n/p$ term. In the n/p term, the $C + 1$ component mixture distribution comprising the Laplace and normal distributions. Specifically, the Laplace distribution with location parameter 0 and scale parameter τ^{-1}, $f_L\left(\hat{\beta}_j; 0, \tau^{-1}\right)$, is assumed as the distribution of the non-outcome-predictive genes considering the above-mentioned feature (ii) of the lasso, while the C-component ($c = 1, \ldots, C$) normal distribution with mean μ_c ($\neq 0$) and variance σ_c^2 is assumed as the distribution of the outcome-predictive genes. It should be noted that normal distribution is a choice of convenience. Next, in the $1-n/p$ term, the Laplace distribution with location parameter 0 and scale parameter ϵ is assumed as the distribution of the $p - n$ genes, considering the above-mentioned feature (i) of the lasso.

Using the estimated mixture distribution, we defined a FPR for a cut-off value ζ (> 0) as follows: given the cut-off value ζ, the area under the Laplace distribution in the n/p term is the estimated proportion of FP genes, and can be written as follows:

$$\begin{aligned}\hat{P}_{\mathrm{FP}} &= \hat{\pi}_0\left[\int_{-\infty}^{-\zeta} f_L\left(u; 0, \hat{\tau}^{-1}\right) du + \int_{\zeta}^{+\infty} f_L\left(u; 0, \hat{\tau}^{-1}\right) du\right] \\ &= 2\hat{\pi}_0 \int_{\zeta}^{+\infty} f_L\left(u; 0, \hat{\tau}^{-1}\right) du \end{aligned} \tag{2.7}$$

Next, the estimated proportion of true positive (TP) genes for the cut-off value ζ is given by the following:

$$\hat{P}_{\mathrm{TP}} = \sum_{c=1}^{C} \hat{\pi}_c \left[\int_{-\infty}^{-\zeta} f_N\left(u; \hat{\mu}_c, \hat{\sigma}_c^2\right) du + \int_{\zeta}^{+\infty} f_N\left(u; \hat{\mu}_c, \hat{\sigma}_c^2\right) du\right] \tag{2.8}$$

Using equation (2.7) and equation (2.8), we obtain the FPR estimator for the cut-off value ζ as follows:

$$\widehat{\mathrm{FPR}}(\zeta) = \frac{\hat{P}_{\mathrm{FP}}}{\hat{P}_{\mathrm{TP}} + \hat{P}_{\mathrm{FP}}} \tag{2.9}$$

Based on the cut-off value ζ used, the estimated proportions of FP and TP genes and the corresponding estimated FPR are found to vary. We determined a cut-off value based on the target FPR specified *in priori*. Specifically, by sequentially changing ζ, we determined the cut-off value that allowed the estimated FPR to be less than or equal to the target FPR. For example, if the target FPR was 0.05, we used the minimum cut-off value that would make the estimated FPR ≤ 0.05.

2.3 Simulation study

Simulation setting

We performed simulation studies to examine the precision of the FPR estimated by the proposed method. In the simulation studies, the number of patients n is set to 200. The number of genes p is set to 1,000, including the p_1 (= 5, 30) outcome-predictive genes, i.e., TP genes. The coefficient for gene j ($j = 1, \ldots, p$) β_j is set to 1.5 for the outcome-predictive genes ($j = 1, \ldots, p_1$) and 0 for the non-outcome-predictive genes ($j = p_1 + 1, \ldots, p$). The number of component C is set to 1 throughout. We may not be able to assume independence among genes, since the expression levels among the outcome-predictive genes are likely to be correlated because of gene co-regulation. It may be reasonable to assume that the expression levels among the non-outcome-predictive genes as well as those between the outcome-predictive genes and the non-outcome-predictive genes are independent.[44] The gene expression levels for patient i, $\mathbf{x}_i$, are generated from the multivariate normal distribution with mean vector 0 and covariance matrix $\boldsymbol{\Sigma}$ with variance 1, so that the correlation among the expression levels of the outcome-predictive genes is 0.0, 0.2, or 0.5, and is constant among the outcome-predictive genes. The survival time for patient i is generated on the basis of the exponential model as follows:

$$t_i = -\log(U)/\exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \tag{2.10}$$

where U is the uniform random variable between 0 and 1.[45] We set λ to 10−30 by 5 in the simulation studies in order to evaluate the precision of the estimated FPR for various values of λ, although the optimal value of λ is determined by cross validation in practice. The value of ζ is defined as the minimum value among $|\hat{\beta}_j|$ ($\neq 0$) ($j = 1, \ldots, p$) in the simulation studies. The average value for true FPR, the estimated numbers of both TP and FP genes, and the estimated FPR in 1,000 simulations are reported.

Simulation results

Table 2.1 shows that the average of the genes with $\hat{\beta}_j \neq 0$ in the lasso, true FPR, and the estimated TP, FP, and FPR for each design parameters in 1,000 simulations. According to Table 2.1, we found that the accuracy of the estimated FPR varied depending on the value of λ. Specifically, the accuracy of the estimated FPR was satisfactory for the values of $\lambda = 10, 15$, and 20, and it was slightly underestimated for the values of $\lambda = 25$ and 30. The number of genes with $\hat{\beta}_j$ was relatively small for the larger value of λ; therefore, the degree of underestimation observed in the simulation studies may be acceptable. For instance, in case of $\rho = 0.0$, $p_1 = 5$, and $\lambda = 30$, the average number of true and estimated FP genes were 1.3 (= 6.5×0.198) and 1.0 (= 6.5×0.146), respectively, and the difference between them was negligibly small in practice. Furthermore, the values of ρ and p_1 did not greatly impact the accuracy of the FPR estimated.

Table 2.1: Accuracy of the FPR estimated using the method proposed in the simulation studies: a number of patients is 200 and a number of genes is 1,000

ρ	p_1	λ	$\#\left\{j;\hat{\beta}_j \neq 0\right\}$	True FPR, %	$\widehat{\text{FPR}}$, %	$\widehat{\text{TP}}$	$\widehat{\text{FP}}$
0	5	10	126.2	96.0	96.0	5.0	121.1
		15	69.2	92.7	92.6	5.1	64.1
		20	31.0	83.5	82.9	5.2	25.8
		25	12.4	57.4	53.4	5.5	6.9
		30	6.5	19.8	14.6	5.4	1.1
	30	10	106.7	71.7	71.4	30.3	76.5
		15	72.1	57.7	56.8	30.6	41.5
		20	57.8	50.0	44.0	32.0	25.8
		25	42.5	44.4	32.5	28.5	14.1
		30	28.2	36.9	27.9	20.2	8.0
0.2	5	10	122.7	95.9	95.9	5.0	117.6
		15	65.4	92.3	92.2	5.1	60.3
		20	27.8	81.5	80.7	5.2	22.5
		25	10.3	48.6	43.8	5.5	4.8
		30	5.7	10.1	6.8	5.2	0.5
	30	10	64.1	52.8	52.0	30.5	33.6
		15	32.1	6.4	5.0	30.4	1.7
		20	30.0	0.1	0.1	30.0	0.0
		25	30.0	0.0	0.0	30.0	0.0
		30	30.0	0.0	0.0	30.0	0.0
0.5	5	10	119.3	95.8	95.8	5.0	114.2
		15	62.5	91.9	91.8	5.1	57.4
		20	25.4	79.7	78.8	5.2	20.2
		25	9.2	42.6	36.4	5.5	3.6
		30	5.4	6.5	3.3	5.2	0.2
	30	10	59.8	49.5	48.5	30.5	29.3
		15	31.1	3.4	2.1	30.4	0.7
		20	30.0	0.0	0.0	30.0	0.0
		25	30.0	0.0	0.0	30.0	0.0
		30	30.0	0.0	0.0	30.0	0.0

2.4 Application

We illustrated the exclusion of the FP genes from the genes selected by the lasso through the application of the proposed method to a real data set comprising the overall survival in 240 DLBCL patients with the expression of 7,399 genes.[22] The survival times were observed in 138 patients, and the censored times, in 102 patients. The median follow-up time was 3.9 years, and the median survival time was 2.8 years.

We divided the 240 patients into 2 groups; the training data comprised 160 patients, and the validation data, 80 patients, as described by Rosenwald et al.[22] We determined that the optimal value of λ was 27 by performing 5-fold cross validation, resulting in the selection of 12 genes as the outcome-predictive genes. Table 2.2 shows the GenBank accession number, description, and coefficient estimate for each of the 12 genes selected by the lasso.

Table 2.2: The GenBank accession numbers, descriptions, and coefficient estimates of 12 genes selected by the lasso

GenBank accession number	Description	$\hat{\beta}$
AA805575	Thyroxine-binding globulin precursor	−0.1039
X00452	Major histocompatibility complex, class II, DQ alpha 1	−0.1026
LC_29222	–	−0.0927
AF044323	COX15 homolog, cytochrome c oxidase assembly protein (yeast)	0.0167
L19872	Hydrocarbon receptor	−0.0078
M20430	Major histocompatibility complex, class II, DR beta 5	−0.0076
K01171	Major histocompatibility complex, class II, DR alpha	−0.0067
X59812 (R92015)	Cytochrome P450, subfamily XXVIIA polypeptide	−0.0028
M63438	Immunoglobulin kappa constant	0.0028
X82240 (AA729003)	T-cell leukemia/lymphoma 1A	−0.0017
X82240 (R97095)	T-cell leukemia/lymphoma 1A	−0.001
X59812 (H98765)	Cytochrome P450, subfamily XXVIIA polypeptide	−0.0002

Given the estimated coefficients $\hat{\beta}_j$ $(j = 1, \ldots, 7,399)$, we assume that the 2 mixture distributions with $C = 1$ and 2, and compared their fitness by using Akaike Information Criterion (AIC)[46]. AIC is the most well known criterion for determining the number of components in the models. As a result, we selected the value of $C = 1$ for simplicity of interpretation, although the AICs for $C = 1$ and 2 were almost same. Thus, we assumed the mixture distribution with $C = 1$, and obtained the following estimated distribution (Figure 2.1):

$$\begin{aligned} f_{\hat{\beta}_j} = \quad & \tfrac{160}{7399}\left\{0.75 f_L\left(\hat{\beta}_j; 0, 0.0053\right) + 0.25 f_N\left(\hat{\beta}_j; -0.10, 0.0064\right)\right\} \\ & + \tfrac{7239}{7399} f_L\left(\hat{\beta}_j; 0, 10^{-8}\right) \end{aligned} \tag{2.11}$$

The mixed proportions of the Laplace and normal distributions in the n/p term were too small; therefore, we enlarged the part including these distributions in Figure 2.1. In addition, according to the estimated mixture distribution, the outcome-predictive genes that increase the risk of death, i.e., genes with $\hat{\beta}_j > 0$, were not found.

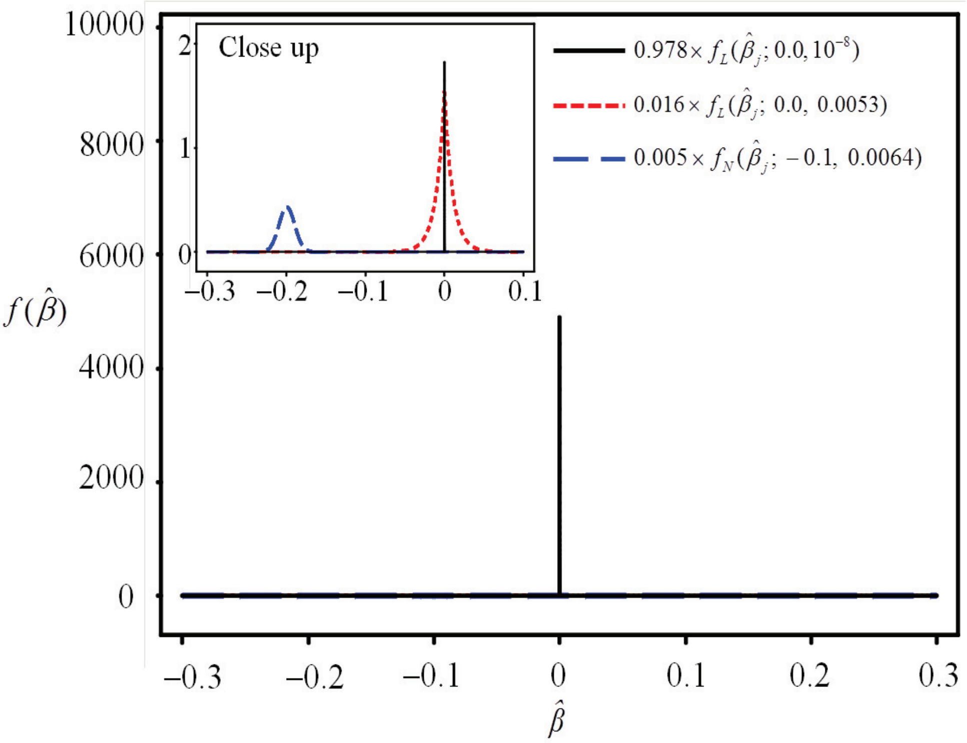

Figure 2.1: The estimated mixture distribution assuming the lasso estimates in the DLBCL data; f_L and f_N are the probability density functions of laplace and normal distributions, respectively. $\hat{\beta}$ is the estimate by the lasso and $f(\hat{\beta})$ is the probability density of $\hat{\beta}$. A magnified image of the distribution between the $\hat{\beta}$ values −0.3 and 0.1 is inserted.

Table 2.3 shows that the estimated numbers of FP and TP genes and the corresponding estimated FPR for various cut-off values. The estimated FPR was less than 5.0% for the cut-off value $\zeta > 0.03$, indicating that 3 genes might be TP genes, although the FPR might be underestimated according to the results of the simulation studies. In order to determine 9 genes that were most likely to be FP genes, we calculated the AICs of all possible models consisting of 3 genes selected among 12 genes, i.e., 220 models in total. The model including 3 genes with $\hat{\beta}$ values of $-0.1039, -0.1026$, and -0.0927 for AA805575, X00452, and LC_29222 showed the lowest AIC, and therefore, the remaining 9 genes were considered as FP genes.

Table 2.3: The estimated numbers of TP and FP genes and the estimated FPR for the cut-off values from 0.0001 to 0.05

cut-off ζ	$\#\left\{j; \lvert\hat{\beta}_j\rvert > \zeta\right\}$	$\widehat{\text{FP}}$	$\widehat{\text{TP}}$	$\widehat{\text{FPR}}$, %
0.0001	12	8.96	3.04	74.6
0.0005	11	8.05	2.95	73.2
0.001	10	7.13	2.87	71.3
0.005	7	3.76	3.24	53.7
0.01	4	1.24	2.76	30.9
0.02	3	0.19	2.81	6.3
0.03	3	0.03	2.97	1.0
0.04	3	0.00	3.00	0.0
0.05	3	0.00	3.00	0.0

Gene Set Enrichment Analysis

As an alternative method for the exclusion of the FP genes from the genes selected by the lasso, we used the Gene Set Enrichment Analysis (GSEA),[47] a computational method that assesses whether an *a priori* defined set of genes shows statistically significant relevance to survival time. The set of genes to be assessed by GSEA is generally defined based on the functional/biological relevance of gene expression profiles, such as genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same Gene Ontology (GO) category. In this study, for the application of the GSEA to the DLBCL data, we identified 1,454 sets of genes based on the GO categories. Of these, 53 gene sets included at least 1 of the 12 genes selected by the lasso method. It should be noted that 5 genes (e.g., M20430, AA805575, M63438, LC_29222, and L19872) were not included in any of the gene sets. For this study, we implemented the modified GSEA for survival time proposed by Lee et al.[48] Table 2.4 shows 38 gene sets with false discovery rate (FDR) < 0.50 estimated by the modified GSEA. According to Table 2.4, the gene sets, including AF044323 and K01171, showed lower *P*-value and FDR, and therefore, we determined these genes as TP genes, and the remaining 10 genes were conveniently considered FP genes.

Table 2.4: Gene sets with FDR < 0.5 in the GSEA

Gene Set	*P*-value	FDR	The genes included in the gene set
BIOSYNTHETIC PROCESS	<0.001	<0.001	AF044323
CELLULAR BIOSYNTHETIC PROCESS	<0.001	<0.001	AF044323
MITOCHONDRIAL PART	0.002	0.035	AF044323
MITOCHONDRION	0.005	0.066	AF044323
MITOCHONDRIAL ENVELOPE	0.008	0.085	AF044323
CYTOPLASMIC PART	0.014	0.093	AF044323, K01171
LYTIC VACUOLE	0.014	0.093	K01171
LYSOSOME	0.014	0.093	K01171
VACUOLE	0.022	0.103	K01171
CELLULAR COMPONENT ASSEMBLY	0.025	0.103	AF044323
PROTEIN METABOLIC PROCESS	0.028	0.103	AF044323
CELLULAR MACROMOLECULE METABOLIC PROCESS	0.028	0.103	AF044323
SECONDARY METABOLIC PROCESS	0.029	0.103	AF044323
PIGMENT BIOSYNTHETIC PROCESS	0.029	0.103	AF044323
PIGMENT METABOLIC PROCESS	0.029	0.103	AF044323
CELLULAR PROTEIN METABOLIC PROCESS	0.034	0.109	AF044323
MITOCHONDRIAL MEMBRANE	0.035	0.109	AF044323
CYTOPLASM	0.039	0.115	AF044323, K01171
HEME BIOSYNTHETIC PROCESS	0.047	0.125	AF044323
HEME METABOLIC PROCESS	0.047	0.125	AF044323
HETEROCYCLE METABOLIC PROCESS	0.067	0.169	AF044323
MACROMOLECULAR COMPLEX ASSEMBLY	0.082	0.198	AF044323
COFACTOR BIOSYNTHETIC PROCESS	0.106	0.244	AF044323
PROTEIN COMPLEX ASSEMBLY	0.111	0.245	AF044323
COFACTOR METABOLIC PROCESS	0.134	0.284	AF044323
MITOCHONDRIAL INNER MEMBRANE	0.143	0.292	AF044323
RECEPTOR ACTIVITY	0.184	0.349	X00452
MULTICELLULAR ORGANISMAL DEVELOPMENT	0.191	0.349	X82240
TRANSMEMBRANE RECEPTOR ACTIVITY	0.191	0.349	X00452
ORGANELLE INNER MEMBRANE	0.200	0.349	AF044323
CELLULAR PROTEIN COMPLEX ASSEMBLY	0.209	0.349	AF044323
ENVELOPE	0.217	0.349	AF044323
ORGANELLE ENVELOPE	0.217	0.349	AF044323
ORGANELLE PART	0.324	0.467	AF044323
INTRACELLULAR ORGANELLE PART	0.324	0.467	AF044323
INORGANIC CATION TRANSMEMBRANE TRANSPORTER ACTIVITY	0.324	0.467	AF044323
MITOCHONDRIAL MEMBRANE PART	0.326	0.467	AF044323
CYTOCHROME C OXIDASE ACTIVITY	0.356	0.497	AF044323

Prediction accuracy

We demonstrated that the 9 genes identified did not impact the survival outcome by comparing the prediction accuracy between the models consisting of the aforementioned 3 and all 12 genes. Furthermore, we also compared the prediction accuracy between the models by which 3 TP genes were identified by the proposed method and 2 TP genes were identified by the GSEA. For the validation data including 80 patients, the following 3 criteria were calculated: *P*-value for the log-rank test, *P*-value for the prognostic index, and deviance. The 80 patients were categorized into 2 groups by the boundary of the median of prognostic index $\hat{\eta}_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$; the "better" and "worse" prognostic groups. The Kaplan-Meier curves between the 2 groups were compared by the log-rank test. Next, we calculated the *P*-value for the parameter α multiplied by the prognostic index $\hat{\eta}_i$ in the Cox proportional hazard model $h(t_i|\mathbf{x}) = h_0(t)\exp(\alpha\hat{\eta}_i)$. Finally, the deviance was calculated by $-2\left\{l^{(validation)}\left(\hat{\beta}_{training}\right) - l^{(validation)}\left(\mathbf{0}\right)\right\}$ where $l^{(validation)}\left(\hat{\beta}_{training}\right)$ and $l^{(validation)}\left(\mathbf{0}\right)$ are the Cox log partial-likelihood function for the estimated coefficients by using training data and zero vector **0**, respectively. For each criterion, the smaller value suggests better prediction accuracy. The values of the 3 indices for the 3 models−the proposed method that identified 3 TP genes, the lasso method that identified 12 genes, and the GSEA that identified 2 TP genes−are shown in Table 2.5. As shown in Table 2.5, the values of the 3 indices between the models that identified 3 and 12 TP genes are almost the same. Furthermore, the prediction accuracy of the proposed method was found to be better than that of the GSEA. Thus, by using the proposed method, we are able to exclude the genes that are not likely to impact the survival outcome.

Table 2.5: Three criteria for model evaluation

Criteria	Model with 3 genes identified by the proposed method	Model with 12 genes	Model with 2 genes identified by the GSEA
P-value of the log-rank test	0.002	0.007	0.246
P-value of the prognostic index	0.002	0.002	0.120
Deviance	−8.942	−9.072	−1.967

2.5 Summary

In this study, we developed a method to estimate FPR by assuming the mixture distribution comprising the Laplace and normal distributions on the lasso estimates. In practice, we identified the outcome-predictive genes by performing the lasso, and subsequently, removing the FP genes using the proposed method.

Although the penalized regression analyses including the lasso are attractive in the high-dimensional gene expression data, it is difficult to identify the outcome-predictive genes without FP genes by using these methods. Utilizing the proposed method, we can validate the results of the lasso, and identify the outcome-predictive genes more precisely. The assumed mixture distribution was formulated considering the 2 features of the lasso, although it may be a “somewhat complex” distribution. The validity of this assumption was demonstrated through the simulation studies. Specifically, the accuracy of the FPR estimated by the proposed method was satisfactory in many cases. The accuracy was slightly decreased for the larger value of tuning parameter λ, but the underestimation of FPR may be acceptable in practice, as discussed in the Simulation section.

Chapter 3

Developing a survival prediction model with enhancing the lasso approach on gene expression data

3.1 Introduction

Here, we proposed an algorithm to sequentially fit the mixture distribution for this solution path, and we used a simulation study to test the precision of the algorithm when estimating the number of TP. We further demonstrated the proposed algorithm using a well-known diffuse large B-cell lymphoma (DLBCL) dataset comprising overall survival of 240 DLBCL patients and gene expression data of 7,399 genes.[22]

3.2 Method

3.2.1 Solution path of the lasso estimates

Goeman[33] introduced a method to calculate the solution path of the lasso estimates as a function of λ, $\hat{\boldsymbol{\beta}}(\lambda)$, which is based on the algorithm developed by Park and Hastie.[49] The method maximizes $l_p(\boldsymbol{\beta}, \lambda)$ at a fixed λ based on a combination of gradient ascent optimization with the Newton-Raphson algorithm. $\hat{\boldsymbol{\beta}}(\lambda)$ are calculated for $\lambda_0 > \ \ldots > \lambda_k > \ \ldots > \lambda_z > 0$ successively, starting from $\lambda_0 = \max_j \partial l / \partial \beta_j |_{\beta_j = 0}$ (which gives $\hat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0}$ because the value has zero gradients). λ_z is chosen arbitrarily, but is often set to $0.05 \times \lambda_0$ in analyses of gene expression data.[50] The lasso estimates at a current step are set to initial values for calculation of the subsequent step. Step length $\Delta_k = \lambda_k - \lambda_{k+1}$ is the minimum decrement to change the number of selected genes $m^{(k)} \left(= \# \left\{ j; \hat{\beta}_j \left(\lambda_k\right) \neq 0 \right\}\right)$, i.e. only one gene is newly selected or excluded from λ_k to λ_{k+1}.

3.2.2 Proposed algorithm for monitoring TP

We propose an algorithm to sequentially fit the mixture distribution in Equation (2.6) to the solution path of the lasso estimates.[2] In this algorithm, we assumed that the number of TP changed when the newly selected or excluded gene from λ_k to λ_{k+1} was truly correlated to survival, based on the maximum log-likelihood of Equation (2.6). First, we approximated $\hat{P}_{FP} \approx \hat{\pi}_0$ and $\hat{P}_{TP} \approx \sum_{c=1}^{C} \hat{\pi}_c$ in Equations (2.7) and (2.8) by assuming a suitably small cut-off value ζ (≈ 0). We then obtained $\hat{\pi}_0 = \widehat{FP}/m$ and $\hat{\pi}_c = \widehat{TP}_c/m$ ($c = 1, \ \ldots, C$) from Equations (2.7) and (2.8), respectively, where $\widehat{TP}_c$ is an estimate of the number of TP in component c. For $k = 0, \ \ldots, z$, the proposed algorithm was as follows.

Step 1

Step 1.1: In this step, we assumed that the newly selected or excluded gene from λ_k to λ_{k+1} was FP. π_0 denotes the proportion of FP, and is set as

$$\pi_0^{(k+1)} = \begin{cases} \frac{\widehat{FP}^{(k)}+1}{m^{(k+1)}} \text{ if } m^{(k+1)} = m^{(k)} + 1 \\ \frac{\widehat{FP}^{(k)}-1}{m^{(k+1)}} \text{ if } m^{(k+1)} = m^{(k)} - 1 \end{cases}$$

For the other components, c ($c = 1, \ldots, C$), set $\pi_c^{(k+1)} = \widehat{TP}_c^{(k)}/m^{(k+1)}$.

Step 1.2: Given $\hat{\beta}(\lambda_{k+1})$ and $\pi_0^{(k+1)}, \ldots, \pi_C^{(k+1)}$, calculate the maximum log-likelihood of Equation (2.6), $L_0^{(k+1)}$.

Step 2

Step 2.1: Set $c = 1$

Step 2.2: In this step, we assumed that the newly selected or excluded gene from λ_k to λ_{k+1} was TP. For the component c, set

$$\pi_c^{(k+1)} = \begin{cases} \frac{\widehat{TP}_c^{(k)}+1}{m^{(k+1)}} \text{ if } m^{(k+1)} = m^{(k)} + 1 \\ \frac{\widehat{TP}_c^{(k)}-1}{m^{(k+1)}} \text{ if } m^{(k+1)} = m^{(k)} - 1 \end{cases}$$

For the other components, set $\pi_0^{(k+1)} = \widehat{FP}^{(k)}/m^{(k+1)}$ and $\pi_d^{(k+1)} = \widehat{TP}_d^{(k)}/m^{(k+1)}$ ($d = 1, \ldots, C; d \neq c$)

Step 2.3: Given $\hat{\beta}(\lambda_{k+1})$ and $\pi_0^{(k+1)}, \ldots, \pi_C^{(k+1)}$, calculate the maximum log-likelihood of Equation (2.6), $L_c^{(k+1)}$.

Step 2.4: Set $c = c + 1$. Repeat Step 2.2 and 2.3 until $c = C$

Step 3: In this step, we determined whether the newly selected or excluded gene from λ_k to λ_{k+1} was TP or FP based on the maximum log-likelihood which was calculated in Step 1.2 and 2.3. If $L_0^{(k+1)}$ was the largest in $L_c^{(k+1)}$ ($c = 0, \ldots, C$), we assumed that the newly selected or excluded gene was FP; if not, we assumed that it was TP. Therefore, calculate $C_{\max} = \underset{c \in \{0,1,\ldots,C\}}{\arg\max}\, L_c^{(k+1)}$. If $C_{\max} = 0$, update $\widehat{FP}^{(k)}$ as follows:

$$\widehat{FP}^{(k+1)} = \begin{cases} \widehat{FP}^{(k)} + 1 \text{ if } m^{(k+1)} = m^{(k)} + 1 \\ \widehat{FP}^{(k)} - 1 \text{ if } m^{(k+1)} = m^{(k)} - 1 \end{cases}$$

If $C_{\max} > 0$, update $\widehat{TP}_{C_{\max}}^{(k)}$ as follows:

$$\widehat{TP}_{C_{\max}}^{(k+1)} = \begin{cases} \widehat{TP}_{C_{\max}}^{(k)} + 1 \text{ if } m^{(k+1)} = m^{(k)} + 1 \\ \widehat{TP}_{C_{\max}}^{(k)} - 1 \text{ if } m^{(k+1)} = m^{(k)} - 1 \end{cases}$$

Here, calculate the estimated TP at $k + 1$ by $\widehat{TP}^{(k+1)} = \sum_{c=1}^{C} \widehat{TP}_c^{(k+1)}$.

3.3 Simulation study

We performed a simulation study to examine the precision of our estimated TP. In this study, the number of patients, n, was set to 200. The number of genes, p, was set to 1,000, which included the p_1 (= 5 or 30) outcome-predictive genes that are randomly chosen from p genes in each simulation. The coefficient for gene j ($j = 1, \ldots, p$), β_j, was set to 1.5 for the p_1 outcome-predictive genes and 0 for the remaining $p - p_1$ none-outcome-predictive genes. We set λ_z to 5, and the number of components, C, to 1 throughout (although C was determined using a model selection criterion in practice). The gene expression levels for patient i, $\mathbf{x}_i$, were generated from the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ so that the variance was 1 and the correlation $\rho(x_{ik}, x_{il}) = 0$ or $0.5^{|k-l|}$.[51] The survival time for patient i was generated based on the exponential model $t_i = -\log(U)/\exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$ where U is the uniform random variable between 0 and 1.[45] In order to evaluate the precision of the estimated TP for various values of λ, we report a number of selected genes, including true TP, and estimated TP and FP, for λ_k ($k = 5, 10, 50, 100, 150$).

Table 3.1: Accuracy of the estimated number of true positives (TP) obtained using the proposed algorithm in the simulation study. Average of a tuning parameter (λ), number of selected genes $\left(\#\left\{j;\hat{\beta}_j(\lambda) \neq 0\right\}\right)$ in the lasso, true number of true positives (True TP), estimated number of TP ($\widehat{TP}$), and false positives ($\widehat{FP}$) are reported at λ_k ($k = 5, 10, 50, 100, 150$) of the solution path.

p_1	ρ	k	λ	$\#\left\{j;\hat{\beta}_j(\lambda) \neq 0\right\}$	True TP	$\widehat{TP}$	$\widehat{FP}$
30	0	5	47.0	5.0	4.4	2.9	2.2
		10	40.8	10.1	8.0	5.8	4.3
		50	22.9	48.6	25.6	28.5	20.1
		100	12.6	86.7	29.9	32.1	54.7
		150	8.6	124.5	30.0	30.7	93.9
	0.5	5	48.6	5.0	4.1	2.8	2.2
		10	42.1	10.0	7.5	5.8	4.2
		50	23.5	48.1	25.2	31.9	16.3
		100	12.4	84.9	29.9	35.3	49.6
		150	8.4	121.2	30.0	31.6	89.6
5	0	5	66.9	5.0	5.0	3.0	2.0
		10	26.3	10.4	5.0	5.2	5.2
		50	17.2	50.1	5.0	5.2	44.9
		100	12.7	93.9	5.0	5.0	88.9
		150	9.8	128.4	5.0	5.0	123.4
	0.5	5	66.8	5.0	5.0	3.0	2.0
		10	26.5	10.3	5.0	5.2	5.1
		50	16.9	49.5	5.0	5.1	44.4
		100	12.4	92.1	5.0	5.0	87.1
		150	9.6	125.2	5.0	5.0	120.2

Table 3.1 shows the average of λ, a number of selected genes, true TP, and estimated TP and FP, through 1,000 repeats. We observed that the precision of estimated TP varied depending on the value of both p_1 and k. When $p_1 = 5$, the precision of the estimates was sufficient for $k = 10$, 50, 100, and 150, while TP was slightly underestimated for $k = 5$. However, when $p_1 = 30$, the precision of the estimates was sufficient for $k = 5$, 10, and 150, while TP was overestimated for $k = 50$ and 100. For example, when $p_1 = 30$, $\rho = 0.5$, and $k = 100$, the average number of true and estimated TP were 29.9 and 35.3, respectively. The values of ρ did not greatly affect the accuracy of the estimated TP.

3.4 Application

To illustrate how our proposed algorithm could be used to determine λ, we applied it to the DLBCL dataset, comprising survival of 240 DLBCL patients and gene expression data from 7,399 genes[22]. In the gene expression data from the 240 patients, we identified 434 genes with complete sets of gene expression values; all other genes had missing expression values, with an average of 24.7 missing values per gene. Here, we used 0.0 as the missing expression value for descriptive purposes. Similar to Rosenwald[22], we divided the data in two: training data consisting of 160 patients and validation data consisting of 80 patients.

For the training data, we obtained the solution path of the lasso estimates, $\hat{\beta}(\lambda_k)$ $(k = 0, 1, \ldots, z)$. $\lambda_0 = 72.5$ was calculated as described in subsection 2.2. We set $\lambda_z = 3.625$ $(= 0.05 \times \lambda_0)$ according to Simon, et al.[50].

We applied our proposed algorithm to the obtained solution path. We assumed three mixture distributions on the lasso estimates with $C = 1$, 2, or 3, and compared their goodness of fit for the $\hat{\beta}(\lambda_k)$ $(k = 0, 1, \ldots, z)$ by the Akaike information criterion (AIC). As a result, we chose $C = 1$ because it had the best AIC for all λ_k $(k = 0, 1, \ldots, z)$.

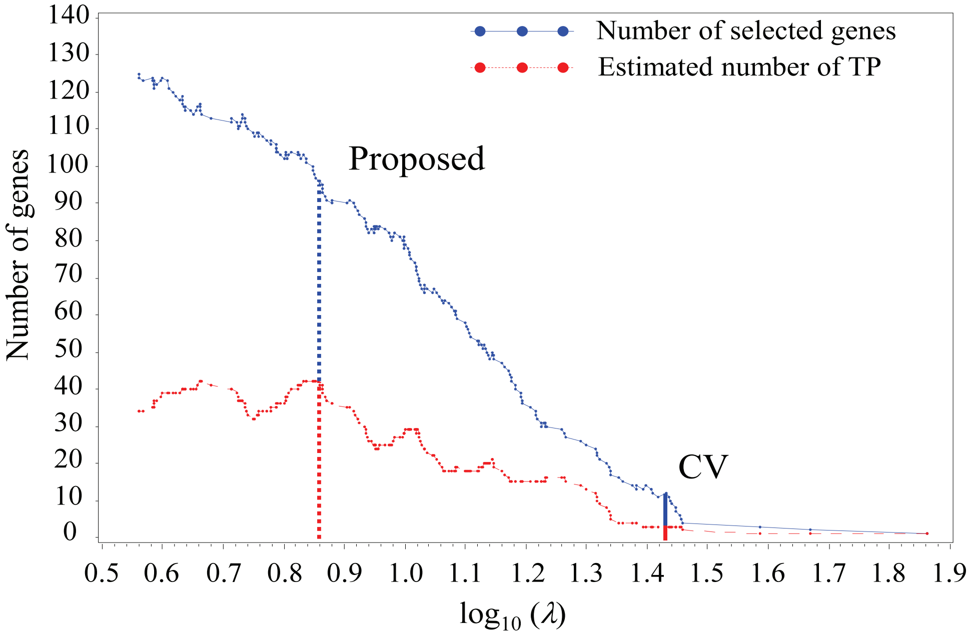

Figure 3.1: Trace plot of number of selected genes and estimated number of true positives (TP) produced by applying the proposed algorithm to the training data from the diffuse large B-cell lymphoma (DLBCL) dataset. We determined $\lambda = 7.19$ ($\log_{10} = 0.86$) as the optimum λ based on the estimated number of TP. Using cross validation (CV), we determined $\lambda = 27$ ($\log_{10} = 1.43$) as the optimum λ.

Figure 3.1 shows the estimated number of TP in a series of values of λ. We found that the lasso selected at most 42 TP, with the number of selected genes at 96, when $\lambda = 7.19$ (= 0.86 as $\log_{10}$). Therefore, we selected $\lambda = 7.19$ as the optimum λ, and the estimated mixture distribution for the value of λ was as follows:

$$f\left(\hat{\beta}_j(7.19)\right) = \frac{160}{7399}\left\{0.57 \times f_L\left(\hat{\beta}_j(7.19); 0, 0.11\right) + 0.43 \times f_N\left(\hat{\beta}_j(7.19); 0.03, 0.11^2\right)\right\} + \frac{7239}{7399} f_L\left(\hat{\beta}_j(7.19); 0, 10^{-8}\right)$$

In order to identify the 42 TP from the 96 selected genes, we arranged the 96 in descending order of $|\hat{\beta}_j|$, and identified the first 42 listed genes with a cut-off value $\zeta = 0.084$. Subsequently, the model that included these 42 genes is identified as the ”42 TP-model”.

In comparison to the 42 TP-model, we performed CV. On the basis of 5-folds CV, 12 genes were selected with $\lambda = 27$ (= 1.43 as $\log_{10}$). Subsequently, the model including these 12 genes is identified as the “CV-model”. Notably, both the 42 TP-model with 42 genes and the CV-model with 12 genes selected 4 genes in common. Table 3.2 shows the GenBank accession number and description for each of the 4 genes selected by both these models.

We compared the prediction accuracy of the 42 TP-model and the CV-model using validation data consisting of 80 patients. For this data, we calculated 3 values that served as comparison criteria: p-values for the log-rank test and prognostic index, and the deviance. The 80 patients were categorized into 2 groups, the “better” and “worse” prognostic groups, using the boundary of the median of prognostic index $\hat{\eta}_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$. The Kaplan-Meier curves between the 2 groups were compared with a log-rank test. Next, we calculated the P-value for the parameter α multiplied by the prognostic index $\hat{\eta}_i$ in the Cox proportional hazard model $h(t_i|\mathbf{x}) = h_0(t)\exp(\alpha\hat{\eta}_i)$. Finally, the deviance was calculated by $-2\left\{l^{(validation)}\left(\hat{\boldsymbol{\beta}}_{training}\right) - l^{(validation)}(\mathbf{0})\right\}$ where $l^{(validation)}\left(\hat{\boldsymbol{\beta}}_{training}\right)$ and $l^{(validation)}(\mathbf{0})$ are the Cox log partial-likelihood function for the estimated coefficients by using the training data and zero vector $\mathbf{0}$, respectively. For each criterion, the lower value suggested better prediction accuracy.

Table 3.3 shows the values of the 3 criteria for each model. We found that the values of all 3 criteria for the 42 TP-model were lower than those for the CV-model, suggesting that the model based on the proposed method was more accurate. Therefore, by using our proposed algorithm, we determined λ and were able to select important genes, likely to be correlated with survival, which the CV was unable to select.

Table 3.2: GenBank accession numbers and descriptions for 4 genes selected by both CV and the model including the 42 genes identified by the algorithm that we developed

GenBank accession number	Description
X82240 (AA729003)	T-cell leukemia/lymphoma 1A
AA805575	Thyroxine-binding globulin precursor
LC_29222	-
X59812(H98765)	Cytochrome P450, subfamily XXVIIA polypeptide

Table 3.3: Values of the comparison criteria for the model including 12 genes determined by CV (CV-model) and the model including the 42 genes identified by our developed algorithm (42 TP-model)

Criteria	CV-model	42 TP-model
P-value of the log-rank test	0.007	<0.001
P-value for the prognostic index	0.002	<0.001
Deviance	−9.079	−11.297

3.5 Summary

In this study, we proposed an algorithm for estimating the number of TP on the solution path of lasso estimates. Monitoring and determining the number of TP for a series of values λ is important because it can increase the probability of uncovering all outcome-predictive genes. The number of TP should be estimated with appropriate accuracy. To confirm the accuracy of our TP, we conducted a simulation study using a typical gene expression dataset. We found that the precision of our algorithm for estimating the number of TP was adequate, although an overestimation occurred with some values of λ. However, the overestimation occurred when the true number of TP was saturated, and so it may not cause a problem by passing over genes that truly correlated with survival.

Chapter 4

An individualized dose-finding approach using the penalized regression for gene mutation patterns in phase I trials for molecularly targeted agents

4.1 Introduction

In this chapter, we propose a novel dose-finding approach to identify the IOD for MTAs in phase I trials in oncology. We utilize the L_1 and L_2 penalized regression, and the IOD determination and gene selection are simultaneously performed based on the elastic net. For each binary efficacy and toxicity outcome, we assume the logistic model with including dose and quadratic form of dose, and gene effect as covariates in main term, and dose by gene effect as covariates in interaction term. The quadratic form of dose in the model is incorporated to account for the non-monotonic patterns for dose-efficacy and dose-toxicity relationships as Cai et al.[42] The estimation of the coefficients for efficacy and toxicity outcome are performed separately and based on the frequentist approach by maximizing the penalized log-likelihood. The dose-finding algorithm is based on the predictive values which are calculated by the estimated penalized regression model. Many dose-finding approaches for MTAs in the available literature take into account correlations between efficacy and toxicity outcomes, and we also consider them based on the multinomial distribution for bivariate joint binary probability of efficacy and toxicity outcome as Sato et al.[36] We implement the simulation studies, and compare the operating characteristics between the proposed method and the method of Wages and Tait[38] with executing the elastic net only once at the end of the trial.

4.2 Method

Notations for data
Let Y_{E_i} and Y_{T_i} denote a binary efficacy and toxicity outcomes for the ith ($i = 1, \dots, n$) entered patient, respectively. Y_{E_i} (or Y_{T_i}) = 1 indicates that efficacy (or toxicity) is observed, and Y_{E_i} (or Y_{T_i}) = 0 indicates otherwise. The gene mutation pattern for ith patient is denoted as $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$, where $x_{ij} = 1$ if jth gene of ith patient has any mutation, $x_{ij} = -1$ otherwise, and p is a total number of genes. The dose for ith patient is denoted as d_i. One dose from dose set $D = (z_1, \dots, z_K)$ where z_k $(k = 1, \dots, K)$ is actual dose is assigned for each patient when entering the trial. In this paper, we define the standardized dose as $z_k^{'} = \log(z_k) - K^{-1} \sum_{l=1}^{K} \log(z_l)$ with the corresponding standardized dose set $D^{'} = (z_1^{'}, \dots, z_K^{'})$ as Sato et al[36], and utilize it as value of dose.

Model 1: L_1 and L_2 penalized regression model separately applied for efficacy and toxicity outcomes based on the binomial distribution

The logistic model is widely used for binary outcomes for evaluating the relationship between efficacy (or toxicity) outcomes and covariates (i.e., gene mutation patterns and dose). The logit of probability to observe efficacy (or toxicity) for ith patient with d_i and x_i is modeled by

$$\text{logit}(Pr(Y_{qi} = 1|d_i, \mathbf{x}_i)) = \eta_q(d_i, \mathbf{x}_i) = b_{0q} + \alpha_{1q}d_i + \alpha_{2q}d_i^2 + \sum_{j=1}^{p}\beta_{jq}x_{ij} + \sum_{j=1}^{p}\gamma_{jq}d_i x_{ij} \; (q = E, T) \quad (4.1)$$

where b_{0q}, $\boldsymbol{\alpha}_q = (\alpha_{1q}, \alpha_{2q})$, $\boldsymbol{\beta}_q = (\beta_{1q}, \ldots, \beta_{pq})$, $\boldsymbol{\gamma}_q = (\gamma_{1q}, \ldots, \gamma_{pq})$ are regression coefficients of the model and a total number of parameters is $2p + 3$ for each value of q. We denote $\eta_{qi} = \eta_q(d_i, \mathbf{x}_i)$ in the following paper. As Cai et al.[42], we assume non-monotonic patterns of dose-toxicity and dose-efficacy relationships by adding quadratic term of dose in an ordinal linear model. In the general setting where the number of patient when the $(n + 1)$ th patient entered exceeds total number of regression coefficients, the regression coefficients can be estimated by maximizing the log likelihood function for the equation (4.1) on the binary distribution as follows:

$$l_{binary}(b_{0q}, \boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q|d_i, \mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{qi}\eta_{qi} - \log\left(1 + \exp(\eta_{qi})\right)\right) \quad (4.2)$$

This cannot work in case of $n < 2p + 3$, and it frequently happens in the Phase I trials.

In the L_1 and L_2 penalized regression model, the regression coefficients are estimated by maximizing the following penalized log-likelihood function:

$$l_{binary,pen}(b_{0q}, \boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q, \lambda|d_i, \mathbf{x}_i) = l_{binary}(b_{0q}, \boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q|d_i, \mathbf{x}_i) - \lambda\rho_c(\boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q) \quad (4.3)$$

where λ is the tuning parameter which determines the amount of shrinkage and $\rho_c(\boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q)$ is the penalty term. There are several approaches for $\rho(\boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q)$. The least absolute shrinkage and selection operator (lasso)[18] has a good property to simultaneously estimate regression coefficients and select important genes, and is frequently used in cancer prediction studies. However, the model we proposed has a problem of multicollinearity due to moderate to high correlations among biomarkers as well as between main and interaction terms, and the lasso may suffer from these problems. The elastic-net[20] can deal with this problem and in this paper, we utilized the elastic net approach, and $\rho_c(\boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q)$ is given by:

$$\begin{aligned}\rho(\boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\gamma}_q) &= \textstyle\sum_{j=1}^{2} w_{q,1,j}\left\{\frac{1}{2}(1-c)\alpha_{jq}^2 + c|\alpha_{jq}|\right\} + \sum_{j=1}^{p} w_{q,2,j}\left\{\frac{1}{2}(1-c)\beta_{jq}^2 + c|\beta_{jq}|\right\} \\ &\quad + \textstyle\sum_{j=1}^{p} w_{q,3,j}\left\{\frac{1}{2}(1-c)\gamma_{jq}^2 + c|\gamma_{jq}|\right\}\end{aligned} \quad (4.4)$$

where c is the mixing-control parameter which controls proportions of penalty of L_1 and L_2 norm, and $w_{q,\cdot,j}$ $(0 \leq w_{q,\cdot,j} \leq 1)$ are separate shrinkage parameters for each parameter of $\boldsymbol{\alpha}_q$, $\boldsymbol{\beta}_q$, and $\boldsymbol{\gamma}_q$. The regression coefficients are estimated by maximizing the penalized log-likelihood function for efficacy and toxicity outcomes separately, and we utilized the coordinate descent algorithm developed by Friedman et al.[41], which is implemented by R package glmnet.

In order to determine the optimal c and λ, the m-fold cross validation is performed.[41] The data for n patients is divided into m-folds. The lth dataset $(l = 1, \ldots, m)$ out of the m divided datasets is used as the training dataset, and the remaining $m - 1$ dataset is used as the test dataset. The deviation of the test dataset is calculated based on coefficients estimated by the training dataset. This step is repeated for m times for all lth $(= 1, \ldots, m)$ datasets. Average values of the deviation is calculated given (c, λ) value. In this paper, $m = n$ is used, which is equal to the leave-one-out cross validation (LOOCV). The LOOCV is repeated for several values of c and λ, and the combination of (c, λ) which gives the best value of the deviation can be determined. In the R package glmnet, λ values where the number of non-zero estimates of coefficients changes in series of values of λ are automatically used for the LOOCV while we have to provide the list of c values. We have to avoid to set $c = 1$, which results in L_1 penalized regression, for dealing with the multicollinearity, and $c = 0$, which results in L_2 penalized regression, because there is no ability to select genes; we used $c = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ throughout this paper.

Fundamentally the objective of the Phase I trial for MTAs is to determine the optimal dose for later phase of developments. Ideally, $w_{q,\ 1,\ 1}$ should be 0 so that the term α_{1q} remains in any cases of (c, λ). Moreover in case that all other estimates than b_{0q} and α_{1q} are zero, the proposed method can seem the ordinal logistic regression model with including only intercept and dose effect. However, in the setting with few patients with several numbers of genes we often face complete or quasi-complete separation problem in logistic regression, and obtain infinite or unrealistically large estimates. Therefore we need a little amount of shrinkage on $w_{q,\ 1,\ 1}$, and it is small enough for estimates to be non-zero. As far as we know, there is no universally acceptable method to determine the value of $w_{q,\ \cdot,\ j}$. In this paper, we set 0.1 for $w_{q,\ \cdot,\ j}$ and 1 for others.

Model 2: The L_1 and L_2 penalized regression model for the multinomial distribution

Let Y_i be a 4 category multinomial variable, and Y_i is derived by Y_{E_i} and Y_{T_i} as $Y_i = 1$ if $Y_{E_i} = 0$ and $Y_{T_i} = 0$, $Y_i = 2$ if $Y_{E_i} = 0$ and $Y_{T_i} = 1$, $Y_i = 3$ if $Y_{E_i} = 1$ and $Y_{T_i} = 0$, and $Y_i = 4$ if $Y_{E_i} = 1$ and $Y_{T_i} = 1$. As in Zhu et al.[52], the Y_i with η_{gi} (g = 1, 2, 3, 4) is modeled by

$$Pr(Y_i = g|d_i, \mathbf{x}_i) = \frac{\exp(\eta_{gi})}{\sum_{l=1}^{4} \exp(\eta_{li})}, \ (g = 1, 2, 3, 4) \tag{4.5}$$

In the general setting, parameters are estimated by maximizing the following log-likelihood function based on the multinomial distribution:

$$l_{multi}(\boldsymbol{b}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|d_i, \mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{g=1}^{4} I(Y_i = g)\eta_{gi} - \log\left(\sum_{l=1}^{4} I(Y_i = l)\eta_{li}\right)\right) \tag{4.6}$$

where $\boldsymbol{b}_0 = (b_{01}, \ldots, b_{04})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_4)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_4)$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_4)$. The penalized log-likelihood function based on the multinomial distribution is given by

$$l_{binary,multi}(\boldsymbol{b}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda|d_i, \mathbf{x}_i) = l_{multi}(\boldsymbol{b}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|d_i, \mathbf{x}_i) - \lambda\rho_c(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \tag{4.7}$$

where like the model 1, ρ_c has a form of the elastic net penalty and given by

$$\begin{aligned} \rho(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \ &= \textstyle\sum_{g=1}^{4}[\sum_{j=1}^{2} w_{g,1,j}\left\{\frac{1}{2}(1-c)\alpha_{jg}^2 + c|\alpha_{jg}|\right\} + \sum_{j=1}^{p} w_{g,2,j}\left\{\frac{1}{2}(1-c)\beta_{jg}^2 + c|\beta_{jg}|\right\} \\ &\quad + \textstyle\sum_{j=1}^{p} w_{g,3,j}\left\{\frac{1}{2}(1-c)\gamma_{jg}^2 + c|\gamma_{jg}|\right\}] \end{aligned} \tag{4.8}$$

The optimal (c, λ) value is determined based on the LOOCV, and $w_{g,\ \cdot,\ \cdot}$ are set in the same way as the Model 1.

Dose-finding algorithm

For estimation in the Model 1, two or more observations for each category (0 or 1) for both toxicity and efficacy outcomes needed for estimating parameters by the coordinate descent method in R package glmnet. We utilize the LOOCV and to keep this condition for all divided datasets, three or more observations for each category (0 or 1) for both toxicity and efficacy outcomes needed. Similarly, for estimation in the Model 2, three or more for each category for both toxicity and efficacy outcomes needed.

Before starting the dose-finding algorithm based on the model 1 and 2, we incorporate the Run-in period at an early stage of the trial wherein the first cohort of patients are treated with the lowest dose and the dose is escalated by the continual reassessment method for the toxicity outcome.[53] In this study, a cohort consisted of three patients. The Run-in period continues while the number of patients is less than N_1 patients, and in this paper we utilize $N_1 = 21$ as Guo et al.[39] After the Run-in period, we initiate the Model-based dose-finding period. If we observe 3 or more observations for each category (0 or 1) for both toxicity and efficacy outcomes, we use Model 1. Nevertheless, if we observe 3 or more observations for each category of Y, we use Model 2. In this paper, we set the cohort size of the model-based dose-finding period to 1 as Guo et al.[39] Note that more number of patients per cohort can also be used; in that case, different doses are assigned to each patient based on the gene mutation patterns.

We define the clinically accepted toxicity probability (ϕ_T). Given the gene mutation patterns of $(n+1)$ th patient, predictive probabilities to observe efficacy and toxicity outcomes which are functions of the given gene mutation patterns and doses are calculated based on the model and estimated parameters. The optimal dose is determined as a dose with a maximal efficacy predictive probability where toxicity predictive probability is equal to or less than ϕ_T, and is assigned to $(n+1)$ th patient. If there is no optimal dose for $(n + 1)$ th patient, then another patient is enrolled. If we cannot enroll patients n_{skip} times in a row, the study is terminated.

We apply this algorithm until the maximal sample size (N_{max}) is reached or the study is terminated. Then we calculate the predictive probabilities for all gene mutation patterns and determine the IOD in the same way as determining the optimal dose in the model-based dose-finding period.

Gene selection method

The model which determined the IOD is also used for performing gene selection. We term the genes which determines the dose-efficacy and dose-toxicity relationships as "DEDT-related genes". The main and interaction terms have different means for the MTAs. The DEDT-related genes in main term are considered as the prognostic biomarker which determine the prognosis of the patient with the given gene mutation pattern while the DEDT-related genes in interaction term are considered as the predictive biomarker which can predict the efficacy and (or) toxicity of dose of the MTAs for the patient with the given gene mutation pattern. Therefore, we perform the selection of genes for main and interaction term separately and both.

If one of $\hat{\beta}_{jq}$ $(q = E,\ T)$ when the model 1 used ($\hat{\beta}_{jg}$ $(g = 1,\ 2,\ 3,\ 4)$ when the model 2 is used) is not equal to 0, the j th gene is selected as the DEDT-related genes as main term. If one of $\hat{\gamma}_{jq}$ $(q = E,\ T)$ ($\hat{\gamma}_{jg}$ $(g = 1, 2, 3, 4)$) is not equal to 0, the j th gene is selected as the DEDT-related genes as interaction term. If one of j th gene of either main or interaction term is selected as the DEDT-related genes, then the j th gene is selected as the DEDT-related genes for either main or interaction term.

4.3 Simulation study

Simulation setting

We implemented simulation studies to compare the accuracy of IOD determination and gene selection of the proposed method with the method proposed by Wages and Tait.[38] We considered the dose set $D = \{1, 2, 3, 4, 5, 6\}$ with six actual doses. Given these actual doses, standardized doses were $D^{'} = \{-1.097,\ -0.403,\ 0.002,\ 0.290,\ 0.513,\ 0.695\}$. The starting dose was set as the lowest dose. We considered the number of genes $p = 4$ or 8. To generate $\mathbf{x}_i$, at first we generated p length vector of random values, $\mathbf{x}^{'}_{i}$, from p multivariate normal distribution $N_p(0,\ \mathbf{\Sigma})$ where $\mathbf{\Sigma} = [\sigma_{kl}]$ $(k,\ l = 1,\ \ldots,\ p)$, and set $\sigma_{kl} = 1$ $(k = l)$, $\sigma_{kl} = 0.5$ $(k \neq l)$. Then x_{ij} was generated by $sign(x^{'}_{ij})$ where $sign(x^{'}_{ij}) = 1$ if $x^{'}_{ij} \geq 0$ and $sign(x^{'}_{ij}) = -1$ otherwise. We simulated and investigated seven different scenarios with respect to the true probabilities of efficacy and toxicity for the standardized doses depending on the gene mutation patterns (Table 4.1). The number of DEDT-related genes for main and interaction term is 1 or 2 depending on the simulation scenario. The DEDT depending on dose and the gene mutation patterns are shown in Figure 4.1 and 4.2. Each simulation consisted of 1,000 trials.

In both methods, we set the maximum sample size N_{max} to 60, and the clinically accepted toxicity probability (ϕ_T) to 0.3. For the proposed method, in Run-in period, we used the (0.01, 0.08, 0.15, 0.22, 0.29, 0.36) for prior toxicity probability to implement the CRM. The n_{skip} is set to 10.

In this study, we utilized the method by Wages and Tait[38] as comparator. In their study, the monotonic dose-toxicity relationship is considered in the setting of MTA developments, and was modeled by using the power model based on one skeleton. We set the skeleton values from the lowest to the highest doses as (0.01, 0.08, 0.15, 0.22, 0.29, 0.36) as Wages and Tait.[38] In contrast, the non-monotonic dose-efficacy relationship is considered and was modeled by using the class of power models including several skeletons that could be seen in MTAs developments as the typical dose-efficacy relationship. We set the eleven sets of skeleton values from the lowest to the highest doses as 1: (0.60, 0.50, 0.40, 0.30, 0.20, 0.10), 2: (0.50, 0.60, 0.50, 0.40, 0.30, 0.20), 3: (0.40, 0.50, 0.60, 0.50, 0.40, 0.30), 4: (0.30, 0.40, 0.50, 0.60, 0.50, 0.40), 5: (0.20, 0.30, 0.40, 0.50, 0.60, 0.50), 6: (0.10, 0.20, 0.30, 0.40, 0.50, 0.60), 7: (0.20, 0.30, 0.40, 0.50, 0.60, 0.60), 8: (0.30, 0.40, 0.50, 0.60, 0.60, 0.60), 9: (0.40, 0.50, 0.60, 0.60, 0.60, 0.60), 10: (0.50, 0.60, 0.60, 0.60, 0.60, 0.60), 11: (0.60, 0.60, 0.60, 0.60, 0.60, 0.60), which were used in the simulation study conducted in the original paper.[38] The sample size for adaptive randomization phase was set to 16. The trial is terminated by safety reason when the lower 95% confidential interval of the posterior probability of the lowest dose exceeds the ϕ_T. We don't utilize the interim trial termination due to futility to be the same condition as the proposed method. It is not comparable with our proposed method to recommend only one dose for overall population estimated by the original method of Wages and Tait.[38] Therefore, as our proposed method, we implemented the elastic net for IOD determination at the end of each trial conducted based on the method of Wages and Tait.[38] The gene selection is also implemented in the same manner as ours. We term this method as the WT-ELNET method.

We compare two kinds of evaluation criteria between the proposed and the WT-ELNET method. One is a correct recommendation rate of IOD for each gene mutation patterns, and also the average of them within all gene mutation patterns. The other is the proportion of correctly (or incorrectly) selected genes among DEDT (or none DEDT) genes, termed as P-CSG (or P-ISG).

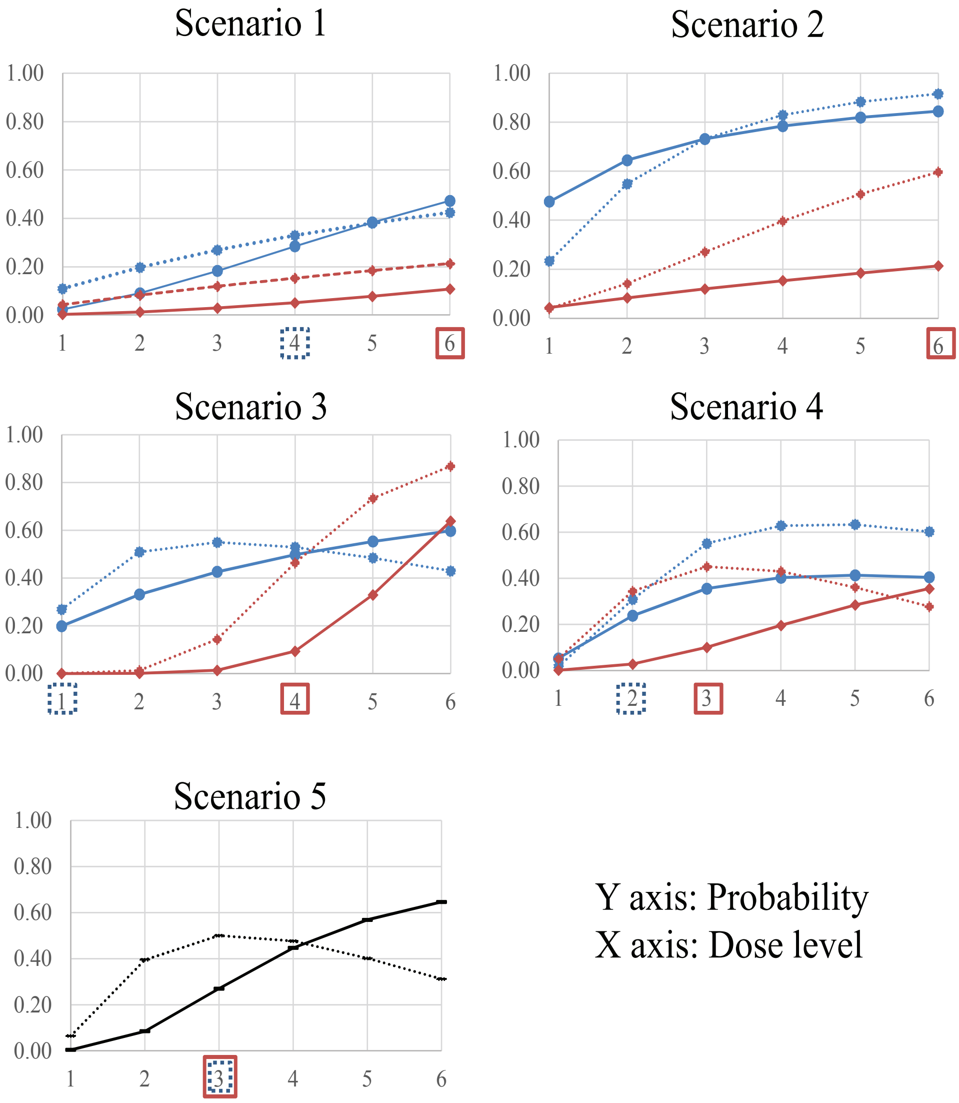

Figure 4.1: Simulation scenarios 1-5: The dotted and solid line are the dose-efficacy and dose-toxicity curves, respectively. A curve with circle and diamond are for $x_1 = 1$ and $x_1 = -1$, respectively. The optimal dose for $x_1 = 1$ and $x_1 = -1$ are indicated by enclosing the dose level in a dotted and solid square, respectively. For scenario 5, the gene does not influence the dose-efficacy and dose-toxicity relationships.

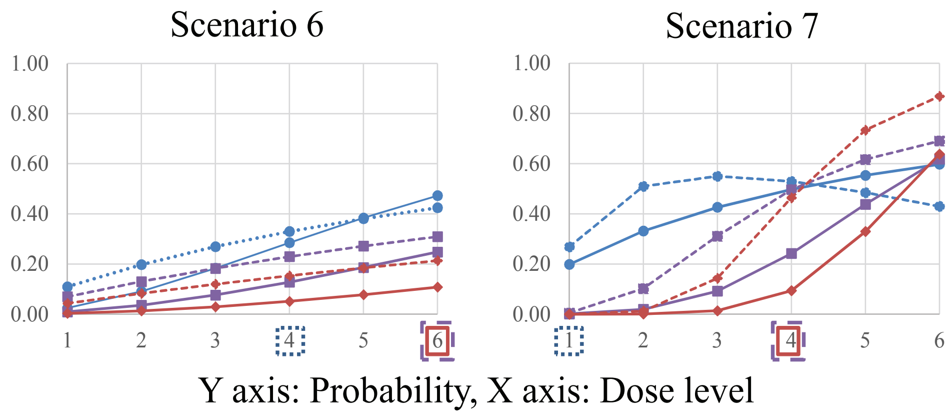

Figure 4.2: Simulation scenarios 6 and 7: The dotted and solid line are the dose-efficacy and dose-toxicity curves, respectively. A curve with circle, square, and diamond are for $(x_1,\ x_2) = (+1, +1)$, $(x_1,\ x_2) = (+1, -1)$ or $(x_1,\ x_2) = (-1, +1)$, and $(x_1,\ x_2) = (-1, -1)$, respectively. The optimal dose for $(x_1,\ x_2) = (+1, +1)$, $(x_1,\ x_2) = (+1, -1)$ or $(x_1,\ x_2) = (-1, +1)$, and $(x_1,\ x_2) = (-1, -1)$ are indicated by enclosing the dose level in a dotted, dashed, and solid square, respectively.

Table 4.1: The data generation model for dose-efficacy and dose-toxicity in each simulation scenario.

Scenario	Model
1	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -2.5 + 2d_i + x_{i1} + u_i$ $\text{logit}(Pr(Y_{Ei} = 1\|d_i, \mathbf{x}_i)) = -1.5 + d_i + 0.5x_{i1} + u_i$
2	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -0.5 + d_i + 1.5x_{i1} + u_i$ $\text{logit}(Pr(Y_{Ei} = 1\|d_i, \mathbf{x}_i)) = 2d_i + x_{i1} + u_i$
3	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -2.3 + 4d_i + 2x_{i1} - 3x_{i1}d_i + u_i$ $\text{logit}(Pr(Y_{Ei} = 1\|d_i, \mathbf{x}_i)) = -0.8 + 3d_i - d_i^2 + x_{i1} - 3x_{i1}d_i + u_i$
4	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -1.4 + 2d_i - d_i^2 + 0.8x_{i1} - 1x_{i1}d_i + u_i$ $\text{logit}(Pr(Y_{Ei} = 1\|d_i, \mathbf{x}_i)) = d_i - 2d_i^2 + 0.2x_{i1} + 0.7x_{i1}d_i + u_i$
5	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -1 + 3d_i - d_i^2 + u_i$ $\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = 0.25d_i - 2d_i^2 + u_i$
6	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -2.5 + 2d_i + 0.5x_{i1} + 0.5x_{i2} + u_i$ $\text{logit}(Pr(Y_{Ei} = 1\|d_i, \mathbf{x}_i)) = -1.5 + d_i + 0.25x_{i1} + 0.25x_{i2} + u_i$
7	$\text{logit}(Pr(Y_{Ti} = 1\|d_i, \mathbf{x}_i)) = -2.3 + 4d_i + x_{i1} + x_{i2} - 1.5x_{i1}d_i - 1.5x_{i2}d_i + u_i$ $\text{logit}(Pr(Y_{Ei} = 1\|d_i, \mathbf{x}_i)) = -0.8 + 3d_i - d_i^2 + 0.5x_{i1} + 0.5x_{i2} - 1.5x_{i1}d_i - 1.5x_{i2}d_i + u_i$

$u_i \sim N(0, 0.5)$ induces correlations between efficacy and toxicity outcomes.

Simulation result

Table 4.2 to 4.5 illustrated simulation results for each simulation scenario in terms of the recommendation rates with p = 4 and 8, respectively. In scenarios, the average of a correct recommendation rate of our proposed method is up to 0.61 and higher compared with the WT-ELNET method. Especially when patients with one gene mutation pattern do not have the IOD while others have (e.g., scenario 2), the proposed method is significantly better than the WT-ELNET in terms of the average of a correct recommendation rate of the IOD. In scenario where genes do not determine the dose-efficacy and dose-toxicity relationships (e.g., scenario 5), it seems that the proposed and WT-ELNET method are comparable. We also calculate the recommendation rate when one recommendation dose is determined by the original WT method, and found that the recommendation rate is 0.72, suggesting that it might be better than the proposed method. However, this values can be obtained only if we absolutely know that no genes affect the dose-efficacy and dose-toxicity relationships and also do not perform the IOD determination, and it may not be comparable between the proposed and the original WT method. It seems that a number of genes which do not influence the dose-efficacy and dose-toxicity relationships does not influence the correct recommendation rate of IOD significantly.

Table 4.6 and 4.7 illustrated operating characteristics of gene selection for each simulation scenario with p = 4 and 8, respectively. When p = 4, The average P-CSG is high up to 0.91 and at least 0.48 in the proposed method, and higher compared with the WT-ELNET method in almost all scenarios. However, in the same time, the relatively high P-ISG were observed up to 0.63 for the proposed method, and equal to or a bit higher than the WT-ELNET method. For each main and interaction term, we found that the P-CSG in main terms is higher than the one in interaction term for both the proposed and the WT-ELNET method. The same results were shown when p = 8, but the values of both P-CSG and P-ISG for all, main, and interaction term a bit decrease compared with p = 4.

Table 4.2: The recommendation rates for the IOD for the proposed and the method proposed proposed by Wages and Tait[38] with implementing the elastic net (WT-ELNET) for scenario 1-5 when $p = 4$

Scenario	The proposed method				The WT-ELNET method			
	Overall	x_1		Early Termination	Overall	x_1		Early Termination
		-1	$+1$			-1	$+1$	
1	0.49	0.74	0.25	0	0.33	0.52	0.13	0
2	0.62	0.42	0.81	0.11	0.42	0.30	0.55	0.03
3	0.37	0.46	0.28	0.01	0.26	0.13	0.38	0
4	0.27	0.28	0.25	0.01	0.24	0.19	0.29	0
5	0.50	NA	NA	0.01	0.32	NA	NA	0

Table 4.3: The recommendation rates for the IOD for the proposed and the method proposed proposed by Wages and Tait[38] with implementing the elastic net (WT-ELNET) for scenario 6 and 7 when $p = 4$

Scenario	The proposed method						The WT-ELNET method					
	Overall	(x_1, x_2)				Early Termination	Overall	(x_1, x_2)				Early Termination
		-1 -1	-1 $+1$	$+1$ -1	$+1$ $+1$			-1 -1	-1 $+1$	$+1$ -1	$+1$ $+1$	
6	0.50	0.73	0.55	0.54	0.19	0	0.36	0.52	0.40	0.42	0.12	0
7	0.33	0.51	0.32	0.32	0.19	0.01	0.21	0.16	0.19	0.20	0.30	0

Table 4.4: The recommendation rates for the IOD for the proposed and the method proposed proposed by Wages and Tait[38] with implementing the elastic net (WT-ELNET) for scenario 1-5 when $p = 8$

Scenario	The proposed method				The WT-ELNET method			
	Overall	x_1		Early Termination	Overall	x_1		Early Termination
		-1	$+1$			-1	$+1$	
1	0.45	0.71	0.20	0	0.31	0.51	0.11	0
2	0.54	0.34	0.74	0.11	0.40	0.31	0.50	0.04
3	0.33	0.42	0.24	0.01	0.25	0.12	0.38	0
4	0.26	0.30	0.23	0	0.22	0.20	0.25	0
5	0.48	NA	NA	0.01	0.33	NA	NA	0

Table 4.5: The recommendation rates for the IOD for the proposed and the method proposed proposed by Wages and Tait[38] with implementing the elastic net (WT-ELNET) for scenario 6 and 7 when $p = 8$

Scenario	The proposed method						The WT-ELNET method					
	Overall	$(x_1,\ x_2)$				Early Termination	Overall	$(x_1,\ x_2)$				Early Termination
		-1	-1	$+1$	$+1$			-1	-1	$+1$	$+1$	
		-1	$+1$	-1	$+1$			-1	$+1$	-1	$+1$	
6	0.50	0.69	0.56	0.55	0.18	0	0.37	0.52	0.43	0.44	0.10	0
7	0.31	0.47	0.31	0.30	0.14	0.01	0.18	0.14	0.17	0.18	0.24	0

Table 4.6: The operating characteristics in terms of gene selection for the proposed and the method proposed proposed by Wages and Tait[38] with implementing the elastic net (WT-ELNET) when $p = 4$

Scenario	The proposed method						The WT-ELNET method					
	P-CSG			P-ISG			P-CSG			P-ISG		
	O	M	I	O	M	I	O	M	I	O	M	I
1	0.90	0.90	NA	0.42	0.56	0.32	0.79	0.79	NA	0.38	0.53	0.27
2	0.92	0.92	NA	0.56	0.59	0.54	0.96	0.96	NA	0.61	0.56	0.66
3	0.73	0.93	0.54	0.50	0.62	0.39	0.77	0.98	0.55	0.57	0.68	0.45
4	0.47	0.67	0.27	0.37	0.50	0.24	0.57	0.84	0.30	0.42	0.57	0.26
5	NA	NA	NA	0.22	0.37	0.08	NA	NA	NA	0.30	0.42	0.17
6	0.71	0.71	NA	0.40	0.54	0.33	0.65	0.65	NA	0.36	0.53	0.27
7	0.64	0.80	0.48	0.52	0.63	0.48	0.73	0.92	0.53	0.61	0.75	0.48

P-CSG (or P-ISG): The proportion of correctly (or incorrectly) selected genes among DEDT (or none DEDT)-related genes, O: Overall, M: Main term, I: Interaction term

Table 4.7: The operating characteristics in terms of gene selection for the proposed and the method proposed proposed by Wages and Tait[38] with implementing the elastic net (WT-ELNET) when $p = 8$

Scenario	The proposed method						The WT-ELNET method					
	P-CSG			P-ISG			P-CSG			P-ISG		
	O	M	I	O	M	I	O	M	I	O	M	I
1	0.85	0.85	NA	0.36	0.50	0.25	0.80	0.80	NA	0.34	0.51	0.20
2	0.91	0.92	NA	0.46	0.50	0.41	0.94	0.94	NA	0.50	0.46	0.53
3	0.68	0.91	0.45	0.38	0.50	0.26	0.72	0.98	0.47	0.44	0.58	0.30
4	0.38	0.57	0.19	0.27	0.39	0.15	0.49	0.80	0.17	0.31	0.49	0.14
5	NA	NA	NA	0.17	0.31	0.04	NA	NA	NA	0.22	0.36	0.09
6	0.65	0.65	NA	0.34	0.47	0.24	0.62	0.62	NA	0.32	0.48	0.20
7	0.56	0.75	0.37	0.40	0.53	0.27	0.63	0.90	0.36	0.47	0.64	0.30

P-CSG (or P-ISG): The proportion of correctly (or incorrectly) selected genes among DEDT (or none DEDT)-related genes, O: Overall, M: Main term, I: Interaction term

4.4 Summary

In this study, we proposed a novel dose-finding approach to find the individualized optimal dose for single MTA in phase I trial. The proposed method utilized the L_1 and L_2 penalized regression model to estimate the parameters of the model and perform gene selection with a large number of covariates of genes in main terms and genes by dose effect in an interaction term. The proposed method selects the individualized optimal dose for future patients based on the predictive probability of the toxicity and efficacy outcome of the estimated penalized regression model. The simulation studies demonstrated that the operating characteristics in the proposed method were more favorable than those of the WT-ELNET method especially when the IOD for the one type of gene mutation pattern does not exist while the IODs for others exist (e.g., scenario 2 in simulation studies). These situations can occur such as gefitinib which shows significantly better progression free survival for non-small cell lung cancer patients with EGFR mutation than the standard treatment while not for patients without EGFR mutation.[54] In such a case, the ordinal dose-finding approach without considering gene mutation patterns does not work well because it does not allow us to increase dose or mistakenly terminates a trial due to high toxicity and/or low efficacy probability observed in only subpopulation with one type of gene mutation pattern. Therefore, our proposed method is useful especially for developments of MTAs.

Chapter 5

Discussion

5.1 Gene selection using a high-dimensional regression model with gene expression data in cancer prognostic studies

In this study, we developed a method to estimate FPR by assuming the mixture distribution comprising the Laplace and normal distributions on the lasso estimates. In practice, we identified the outcome-predictive genes by performing the lasso, and subsequently, removing the FP genes using the proposed method.

Although the penalized regression analyses including the lasso are attractive in the high-dimensional gene expression data, it is difficult to identify the outcome-predictive genes without FP genes by using these methods. Utilizing the proposed method, we can validate the results of the lasso, and identify the outcome-predictive genes more precisely. The assumed mixture distribution was formulated considering the 2 features of the lasso, although it may be a "somewhat complex" distribution. The validity of this assumption was demonstrated through the simulation studies. Specifically, the accuracy of the FPR estimated by the proposed method was satisfactory in many cases. The accuracy was slightly decreased for the larger value of tuning parameter λ, but the underestimation of FPR may be acceptable in practice, as discussed in the Simulation section.

In the section on Application to the DLBCL Data, the utility of the proposed method was illustrated. We were able to eliminate the FP genes from the genes selected by the lasso with $\lambda = 27$, and improved the accuracy of prediction of the model. We further identified the TP genes and examined the prediction accuracy of overall survival based on them, using the proposed method and GSEA. Both methods identified no TP genes in common. The prediction accuracy using the 3 genes identified by the proposed method outperformed that using the 2 genes identified by the GSEA. The GSEA introduced by the Subramanian et al.[47] evaluates gene expression data at the level of gene sets. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments. In contrast, the proposed method evaluates gene expression data at the level of genes and does not use prior biological knowledge when identifying the outcome-predictive genes.

Some variants of the lasso and penalized regression methods are used, e.g., smoothly clipped absolute deviation penalty (SCAD)[55], adaptive lasso[56,57], elastic net[20], and ridge regression[17], but of these, we chose the lasso in this study, because of our concerns regarding the high possibility of missing the true positives for the SCAD and adaptive Lasso, the difficulty in choosing 2 penalties for the elastic net, and the absence of any property to select genes for ridge regression.[29]

The determination of the value of the tuning parameter λ is required when performing the lasso. The value of λ is frequently determined on the basis of the cross validation that evaluates the adequacy of the model, as explained in the Methods section. By utilizing the proposed method, we could also determine the value of λ by considering not only the prediction accuracy but also the FPR.

5.2 Developing a survival prediction model with enhancing the lasso approach on gene expression data

In this study, we proposed an algorithm for estimating the number of TP on the solution path of lasso estimates. Monitoring and determining the number of TP for a series of values λ is important because it can increase the probability of uncovering all outcome-predictive genes. The number of TP should be estimated with appropriate accuracy. To confirm the accuracy of our TP, we conducted a simulation study using a typical gene expression dataset. We found that the precision of our algorithm for estimating the number of TP was adequate, although an overestimation occurred with some values of λ. However, the overestimation occurred when the true number of TP was saturated, and so it may not cause a problem by passing over genes that truly correlated with survival. In the simulation study where $p_1 = 30$ and $\rho = 0.5$, the maximum average estimated number of TP was 35.3 at $\lambda = 12.4$. Using this λ to select TP, an average selection of 29.9 TP within 30 outcome-predictive genes can be made, with the number of TP genes that are passed over being negligible in practice.

The data that have been provided in Table 2 showed that the number of false positives increased, while the number of true positives increased and then plateaued as the tuning parameter decreased. To decrease the number of FP identified while maintaining an adequate number of TP, we should determine the value of λ by monitoring both the number of TP and the false positive rate (= FP/(TP + FP)) in the proposed method.

Additionally, our proposed algorithm was applied to DLBCL data. We determined the value of the tuning parameter based on the maximum number of estimated TP uncovered by the algorithm. We identified 42 TP genes among 96 selected genes based on the ranking of the absolute values of the lasso estimates. We can also identify TP based on model evaluation criteria such as AIC among all possible combinations of 42 genes from 96, i.e., ${}_{96}C_{42}$ ($> 10^{27}$) combinations in total; however, calculation of AIC for all possible gene combinations is a distant approach. To evaluate the efficiency of the approach using the ranking of the lasso estimates, we calculated the AIC for 10,000 randomly chosen models among all the possible models and subsequently compared it with the AIC of our approach. From 10,000 models, the AIC of 425 models (4.25 %) was better than that of our approach. This result indicated that our ranking-based approach has a satisfactory performance in practice with respect to the identification of 42 genes. Although investigation of all possible gene combinations is ideal, our approach is a good alternative.

In the application to DLBCL data, in comparison to a CV method by which 12 genes were identified, we identified 42 TP genes with our algorithm, and we improved the prediction accuracy of the model. In practice, some researchers might be satisfied with identifying a few promising genes, and would not be unduly worried about passing over others. In such a situation, the CV would be preferable because it developed the model to uncover a few genes with just a small loss of prediction accuracy. However, genes that are selected by the lasso are often investigated with greater scrutiny by genetic researchers, and so passing over outcome-predictive genes by the lasso could represent a major problem. Indeed, if the lasso passes over outcome-predictive genes, some genetic research may not take place. Therefore, when identifying all outcome-predictive genes is a priority, our proposed algorithm will be most useful.

5.3 An individualized dose-finding approach using the penalized regression for gene mutation patterns in phase I trials for molecularly targeted agents

We proposed a novel dose-finding approach to find the individualized optimal dose for a MTA in phase I trial. The proposed method utilized the L_1 and L_2 penalized regression model to estimate the parameters of the model and perform gene selection with a large number of covariates of genes in main terms and genes by dose effect in an interaction term. The proposed method selects the individualized optimal dose for future patients based on the predictive probability of the toxicity and efficacy outcome of the estimated penalized regression model.

The simulation studies demonstrated that the operating characteristics in the proposed method were more favorable than those of the WT-ELNET method especially when the IOD for the one type of gene mutation pattern does not exist while the IODs for others exist (e.g., scenario 2 in simulation studies). These situations could occur such as gefitinib which shows significantly better progression free survival for non-small cell lung cancer patients with EGFR mutation than the standard treatment while not for patients without EGFR mutation.[54] In such a case, the ordinal dose-finding approach without considering gene mutation patterns does not work well because it does not allow us to increase dose for right patients to the MTA treatment or mistakenly terminates a trial due to high toxicity and/or low efficacy probability observed in only subpopulation with one type of gene mutation pattern. And also, if we cannot observe the data for higher dose for subgroup of patients who show the high efficacy at higher doses, then we cannot advance the developments of the novel and evolutional drugs for the subgroup of patients. Simulation results demonstrated that our proposed method works well in such a situation, therefore it is useful especially for developments of MTAs whose dose-efficacy and dose-toxicity relationships would differ depending on the gene mutation patterns.

The favorable point of the proposed method is that we can perform both the IOD determination and DEDT-related gene selection without observing several gene mutation patterns. In recent years, our knowledge for genome in oncology much is advancing, and we found many genes related to a mechanism of cancer. It is not difficult to anticipate that in near future, when we start developments for a novel and evolutional drug for cancer, we have known tens or hundreds of genes which may influence the dose-efficacy and dose-toxicity relationships, and we may have to consider them in the phase I trials. The simulation results with $p = 8$ (256 gene mutation patterns) illustrated that our proposed method perform well although some gene mutation patterns were not observed with limited number of patients (60 patients). The L_1 and L_2 penalized regression model allows us to include any number of p into the regression model regardless of size of n. It is expected that the proposed method might work when the number of genes increase more than 8, but we have to investigate the operating characteristics further when using it in practice.

The proposed method can identify DEDT-related genes with high P-CSG, but high P-ISG was observed. In low sample size and high dimensional setting, it is known that the elastic net possibly include false positive genes as well as true positive genes in the model. We thought that in early phase of development, any risks to pass over the promising DEDT-related genes should be low because if we passed over them then the further research for DEDT-related genes as well as MTAs would not occur and miss the opportunities of developing the efficacious drug for patients in specific subpopulation with several gene mutation patterns. We believe that the high P-ISG is acceptable because P-CSG is also high enough to proceed the developments of MTAs.

Our proposed method can be extended so that it can account for the other types of efficacy and toxicity outcomes as well as biomarkers. For example, the use of the proportional odds model for grading toxicity outcome, instead of the ordinary logistic model we used, could be considered. However, this may cause the problem in parameter estimations because the number of parameters increase as the number of grade increase. Operating characteristics should be further investigated. For the covariates, the use of the gene expression data as continuous variable, instead of the gene mutation status as categorical variable we used, can be considered. It may be useful when the over or under expression of genes could be predictive for the treatment effect such as the trastuzumab for breast cancer patients with over expression of the HER2 gene.[10] Other high-dimensional analysis techniques can also be used. The extension of the proposed method using the mixed penalized regression model for correlated efficacy and toxicity outcomes within a patient is one of the most promising ones, and should be investigated further.

Chapter 6

Conclusion

The issues of this study are concluded as follows:

6.1 Issue 1

The lasso allows us to efficiently select the outcome-predictive genes in the high-dimensional gene expression data, but the difficulty lies in the inclusion of the FP genes among the selected genes. We proposed the mixture distribution for the lasso estimates. The use of the proposed method allows us to eliminate these genes and improve the prediction accuracy of the Cox model. In addition, based on the proposed mixture distribution, we developed a method for estimating the number of true positives for a series of values of a tuning parameter in the lasso. We demonstrated the utility of the developed method through a simulation study and an application to a real dataset. Our results indicated that our developed method was useful for determining a value for the tuning parameter in the lasso, and reducing the probability of passing over genes that are truly correlated with survival.

6.2 Issue 2

We proposed a novel dose-finding approach to find the individualized optimal dose for single MTA in phase I trial. The proposed method utilized the L_1 and L_2 penalized regression model to estimate the parameters of the model and perform gene selection with a large number of covariates of genes in main terms and genes by dose effect in an interaction term. Simulation results illustrated that our proposed design has good operating characteristics, and indicated that our proposed method was useful for the MTAs development when the genes could be related to the dose-efficacy and dose-toxicity relationships.

The results of this study contribute to establishing the precision medicine through efficient use of the penalized regression model in cancer clinical researches.

Acknowledgments

I extremely grateful to Professor Chikuma Hamada who, in addition to his constant guidance and advice, also provided a stimulating research environment and countless opportunities. His kind comments and suggestions concerning matters of a personal nature are also appreciated. Without his encouragement and help, this study would never have been completed.

I am immeasurably indebted to Dr. Akihiro Hirakawa of Nagoya University Graduate School of Medicine, who consistently supported and encouraged me throughout this work. They provided numerous comments and insights each of the many times that they kindly reviewed manuscripts. I learned a lot about scientific research from them.

I also express my appreciation to Professor Hiroyuki Yashima, Professor Hitoshi Watanabe, Professor Tohru Ikeguchi, and Associate Professor Takashi Sozu of Tokyo University of Science, and Dr.Masayuki Ohkura of Phizer Japan Inc. All were members of my doctoral committee and provided many helpful comments and suggestions.

Finally, I express my gratitude to my family, for always supporting me in all my endeavors.

March 2017

Appendix A

An additional simulation study

A.1 Simulation study for investigating the operating characteristics of the lasso for Chapter 1

To further investigate the operating characteristics of the lasso regarding the number of true positive (TP) and false positive (FP), we conducted simulation studies with assuming typical analyses of gene expression data. The objective of this simulation study is to illustrate the number of TP and FP when we utilize the lasso and determine the optimal value of the tuning parameter by the cross validation approach.

A.1.1 Simulation setting

In the typical analyses for gene expression data, the number of patients (n) is at most 200 while the number of genes (p) is more than 1,000. In this simulation study, we set $n = 100,\ 200$ and $p = 1,000,\ 5,000$. For the number of outcome-predictive genes (p_1), we set $p_1 = 5,\ 30$. The true values of regression coefficients (i.e., the amount of effect to survival time of the outcome-predictive genes) are $\beta_j = 1.5$ ($j = 1, 2, \cdots, p_1$), $\beta_j = 0$ ($j = p_1 + 1, \cdots, p$) as Benner et al.[29] The gene expression levels for patient i, $\mathbf{x}_i$, are generated from the multivariate normal distribution with mean vector 0 and covariance matrix $\boldsymbol{\Sigma}$ with variance 1, so that the correlation among the expression levels of the outcome-predictive genes is 0.0, 0.2, or 0.5, and is constant among the outcome-predictive genes. The survival time for patient i is generated on the basis of the exponential model as follows:

$$t_i = -\log(U)/\exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \tag{A.1}$$

where U is the uniform random variable between 0 and 1.[45] We determine the value of λ by utilizing the cross validation approach, and implement the lasso. The average value for true FPR, the estimated numbers of both TP and FP genes, and the estimated FPR in 1,000 simulations are reported.

Table A.1: Simulation results

p	ρ	p_1	n	$\#\left\{j; \hat{\beta}_j \neq 0\right\}$	TP[1]	FP[2]	FPR[3]	λ
1000	0	5	100	43.7	5.0	38.7	88.3	10.5
			200	71.0	5.0	66.0	92.8	14.8
		30	100	13.0	4.9	8.1	58.9	24.4
			200	76.7	27.4	49.3	60.4	16.5
	0.2	5	100	47.5	5.0	42.5	89.2	9.4
			200	72.9	5.0	67.9	93.0	14.3
		30	100	65.1	30.0	35.1	53.4	5.0
			200	142.2	30.0	112.2	78.8	5.2
	0.5	5	100	49.6	5.0	44.6	89.7	8.8
			200	75.1	5.0	70.1	93.2	13.7
		30	100	69.1	30.0	39.1	56.4	3.9
			200	141.7	30.0	111.7	78.8	4.7
5000	0	5	100	42.8	5.0	37.8	87.5	14.0
			200	85.6	5.0	80.6	94.1	17.5
		30	100	8.8	1.6	7.1	78.9	30.4
			200	34.0	11.8	22.3	63.0	34.5
	0.2	5	100	52.9	5.0	47.9	90.4	11.0
			200	90.2	5.0	85.2	94.4	16.5
		30	100	59.6	29.9	29.7	48.7	7.7
			200	155.6	30.0	125.6	80.7	5.5
	0.5	5	100	56.4	5.0	51.4	91.0	9.8
			200	94.5	5.0	89.5	94.6	15.6
		30	100	79.5	30.0	49.5	59.1	4.4
			200	160.1	30.0	130.1	81.0	4.7

1. Number of selected outcome-predictive genes, True positive
2. Number of selected outcome-predictive genes, False positive
3. False positive rate

A.1.2 Simulation results

In the lasso, the number of selected genes decrease as well as the number of TP and FP when the value of λ become larger. The CV determine the optimal value of λ to select the most predictive model by considering trade-off of the number of TP and FP. That is, the CV tries to make the value of λ larger to decrease the number of FP as well make the value of λ smaller to increase the number of TP.

Table A.1 illustrated that in almost all cases the model determined by the CV contain some additional genes (i.e., FP), and at most 141.7 genes averagely. We also found that the number of TP sometimes is not equal to the true number of the outcome-predictive genes, and suggesting potentially miss them especially in small sample size setting (e.g., only 4.9 genes out of 30 outcome-predictive genes were selected when $p = 1,000$, $\rho = 0.0$, $n = 100$; only 1.6 genes out of 30 outcome-predictive genes when $p = 5,000$, $\rho = 0.0$, and $n = 100$). Regarding the value of λ, we found that it varies depending on the simulation scenario. However, the average FPR is consistently high through all simulation scenarios.

A.2 Additional simulation results for Chapter 2

The simulation results in Chapter 2 illustrated that our proposed mixture distribution can estimate the false positive rates well. We set the number of patients n is set to 200 and the number of genes p is set to 1,000. In a typical gene expression data analyses, these numbers may differ depending on the datasets and interest of researches. To investigate the operating characteristics of our proposed method further, we also performed this additional simulation studies when a number of patients and genes differ. We set $n = 100$, and $p = 1,000$ and 5,000. We assume the same settings as chapter 2 for other parameters.

Table A.2 to A.4 illustrated that our proposed method can estimate FPR well in various settings of n and p. This results support our findings in simulation studies in chapter 2.

Table A.2: Accuracy of the FPR estimated using the method proposed in the simulation studies: a number of patients is 100 and a number of genes is 1,000

ρ	p_1	λ	$\#\left\{j;\hat{\beta}_j \neq 0\right\}$	FPR	F$\hat{\text{P}}$R
0	5	10	47.4	89.3	89.1
		15	18.8	72.1	70.0
		20	8.2	34.7	27.3
		25	5.9	13.0	8.1
		30	4.9	6.1	5.7
	30	10	65.4	77.8	62.3
		15	39.5	71.8	57.8
		20	20.8	64.7	55.6
		25	9.2	55.6	48.3
		30	3.4	43.7	28.3
0.2	5	10	42.6	88.1	87.8
		15	14.1	62.8	59.4
		20	5.9	13.3	8.0
		25	5.1	1.1	0.5
		30	5.0	0.1	0.2
	30	10	34.9	13.6	11.3
		15	29.9	2.4	1.4
		20	26.2	0.6	1.7
		25	21.6	0.2	3.1
		30	17.1	0.1	4.7
0.5	5	10	39.7	87.2	86.9
		15	11.7	55.1	50.6
		20	5.3	5.5	2.8
		25	5.0	0.2	0.1
		30	5.0	0.0	0.0
	30	10	31.5	4.6	3.1
		15	29.5	0.1	0.3
		20	27.7	0.0	0.8
		25	24.9	0.0	1.7
		30	21.7	0.0	2.8

Table A.3: Accuracy of the FPR estimated using the method proposed in the simulation studies: a number of patients is 100 and a number of genes is 5,000

ρ	p_1	λ	$\#\left\{j; \hat{\beta}_j \neq 0\right\}$	FPR	$\hat{\text{FPR}}$
0	5	10	66.0	92.4	92.3
		15	35.6	85.6	85.0
		20	15.9	65.9	62.4
		25	8.7	37.6	30.6
		30	6.1	22.7	17.4
	30	10	79.4	90.9	72.9
		15	56.9	89.1	71.8
		20	34.8	86.7	70.4
		25	17.6	83.1	68.6
		30	7.0	77.6	60.4
0.2	5	10	59.4	91.5	91.3
		15	26.4	80.4	79.5
		20	8.7	38.2	31.6
		25	5.4	6.2	3.3
		30	5.1	1.0	0.5
	30	10	44.7	32.2	30.1
		15	32.1	9.1	6.4
		20	26.8	3.0	2.9
		25	21.6	1.0	3.2
		30	17.0	0.4	5.0
0.5	5	10	55.7	91.0	90.8
		15	21.8	76.2	74.8
		20	6.6	20.9	14.6
		25	5.1	0.9	0.4
		30	5.0	0.0	0.0
	30	10	35.1	14.1	12.0
		15	29.6	0.3	0.5
		20	27.8	0.0	0.8
		25	24.9	0.0	1.6
		30	21.6	0.0	3.0

Table A.4: Accuracy of the FPR estimated using the method proposed in the simulation studies: a number of patients is 200 and a number of genes is 5,000

ρ	p_1	λ	$\#\left\{j;\hat{\beta}_j \neq 0\right\}$	FPR	$\hat{\text{FPR}}$
0	5	10	158.6	96.8	96.8
		15	111.5	95.5	95.5
		20	63.3	92.0	91.9
		25	28.2	81.7	80.9
		30	10.9	51.3	46.5
	30	10	169.7	85.0	78.8
		15	137.2	82.2	75.4
		20	102.9	78.7	70.8
		25	72.2	74.4	64.8
		30	47.7	69.2	59.9
0.2	5	10	154.7	96.8	96.7
		15	105.4	95.2	95.2
		20	56.2	91.0	90.8
		25	22.4	76.8	75.6
		30	8.0	33.9	27.1
	30	10	94.0	67.9	67.5
		15	38.7	21.8	19.8
		20	30.3	0.8	0.4
		25	30.0	0.0	0.0
		30	30.0	0.0	0.0
0.5	5	10	151.2	96.7	96.7
		15	101.4	95.1	95.0
		20	51.8	90.2	90.0
		25	19.1	72.7	71.1
		30	6.8	23.4	16.7
	30	10	86.3	65.0	64.5
		15	34.7	13.0	10.9
		20	30.0	0.1	0.0
		25	30.0	0.0	0.0
		30	30.0	0.0	0.0

Bibliography

[1] National Institutes of Health. About the Precision Medicine Initiative Cohort Program, 2016 (https://www.nih.gov/precision-medicine-initiative-cohort-program). Access date:2nd October, 2016.

[2] Kaneko S, Hirakawa A, Hamada C. Gene selection using a high-dimensional regression model with microarrays in cancer prognostic studies. *Cancer Inform*, 2012;11:29−39.

[3] Kaneko S, Hirakawa A, Hamada C. Enhancing the lasso approach for developing a survival prediction model based on gene expression data. *Comput Math Methods Med*, 2015; Article ID 259474.

[4] Kaneko S, Hirakawa A, Hamada C. An individualized dose-finding approach using the penalized regression for gene mutation patterns in phase I trials for molecularly targeted agents. (In preparation)

[5] American cancer society. Cancer facts and figures 2016. 'http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2016/'. Access date: 6th November, 2016.

[6] Foundation for promotion of cancer research. Cancer statistics in Japan '15. 'http://ganjoho.jp/data/reg_stat/statistics/brochure/2015/cancer_statistics_2015.pdf'. Access date: 6th November, 2016.

[7] Ma J, Hobbs BP, Stingo FC. Statistical methods for establishing personalized treatment rules in oncology. *Biomed Res Int*. 2015;2015:1−13.

[8] DiMasi JA, Reichert JM, Feldman L, et al. Clinical approval success rates for investigational cancer drugs. *Clin Pharmacol Ther*. 2013;94(3):329−335.

[9] Hay M, Thomas DW, Craighead JL, et al. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014;32(1):40−51.

[10] Goldhirsch A, Gelber RD, Piccart-Gebhart MJ, et al. 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial. *Lancet*. 2013: 21;382(9897):1021−8.

[11] Albain KS, Barlow WE, Shak S, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol*. 2010;11(1):55−65.

[12] Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Per Med*. 2010;7(1):33−47.

[13] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning, data mining, inference, and prediction. New York: Springer-Verlag; 2001.

[14] Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*. 2004;2:511−522.

[15] Bair E, Hastie T, Paul D, et al. Prediction by supervised principal components. *J. Am. Stat. Assoc*. 2006;101:119−137.

[16] Nygård S, Borgan Ø, Lingjærde OC. Partial least squares Cox regression on genomic data handling additional covariates. Statistical Research Report 5/2006. Department of Mathematics, University of Oslo; 2006. 'http://www.math.uio.no/eprint/stat_report/2006/05-06.html'.

[17] Hoerl AE, Kennard RW. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*. 1970;35:109−47.

[18] Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc, Series B Stat Methodol*. 1996;58:267−88.

[19] Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385−95.

[20] Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc, Series B Stat Methodol*. 2005;67:301−20.

[21] Bøvelstad HM, Nygård S, Størvold HL, et al. Predicting survival from microarray data-a comparative study. *Bioinformatics*. 2007;23:2080−7.

[22] Rosenwald M, Wright G, Chan CW, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*. 2002;346:1937−47.

[23] van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530−6.

[24] Van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347: 1999−2009.

[25] Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671−9.

[26] Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*. 2007;99:147−57.

[27] Cox DR. Regression models and life-table (with disscusion). *J R Stat Soc, Series B Stat Methodol*. 1972;74:187−220.

[28] Ein-Dor L, Kela I, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005;21:178−8.

[29] Benner A, Zuchnick M, Hielscher T, Ittrich C, Mansmann U. High-dimensional Cox model: The choice of penalty as part of the model building process. *Biom J*. 2010;52:50−69.

[30] Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*. 2005;21:3001−8.

[31] van Wieringen WN, Kun D, Hampel R, Boulesteix AL. Survival prediction using gene expression data: a review and comparison. *Comput Stat Data Anal*. 2009;53:1590−1603.

[32] Verweij PJ, Houwelingen HC. Cross-validation in survival analysis. *Stat Med*. 1993;12:2305−14.

[33] Goeman J. L1 penalized estimation in the Cox proportional hazards model. *Biom J*. 2010;52:70−84.

[34] Thall PF, Cook JD. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 2004; 60: 684−693.

[35] Hirakawa A. An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Stat Med*. 2012; 31: 516−532.

[36] Sato H, Hirakawa A, Hamada C. An adaptive dose-finding method using a change-point model for molecularly targeted agents in phase I trials.*Stat Med*. 2016; 35: 4093−109.

[37] Riviere MK, Yuan Y, Dubois F, Zohar S. A Bayesian dose finding design for clinical trials combining a cytotoxic agent with a molecularly targeted agent. *J R Stat Soc, Series C Appl Stat*. 2015; 64: 215−229.

[38] Wages NA, Tait C. Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *J Biopharm Stat*. 2015; 25(5):903−920.

[39] Guo B, Yuan Y. Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *J R Stat Soc, Series C Appl Stat*. (Accepted).

[40] Reis-Filho JS. Next-generation sequencing. *Breast Cancer Research*. 2009; 11(Suppl 3): S12.

[41] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010; 33(1):1−22.

[42] Cai C, Yuan Y, Ji Y. A Bayesian dose finding design for oncology clinical trials of combinational biological agents. *J R Stat Soc, Series C Appl Stat*. 2014; 63(1):159−173.

[43] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32,407−51.

[44] Tsai CA, Wang SJ, Chen DT, Chen JJ. Sample size for gene expression microarray experiments. *Bioinformatics*. 2005;21:1502−8.

[45] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2006;24:1713−23.

[46] Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19:716−23.

[47] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102:15545−50.

[48] Lee S, Kim J, Lee S. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics*. 2011;12:377.50.

[49] Park MY, Hastie T. L_1-regularization path algorithm for generalized linear models. *J R Stat Soc, Series B Stat Methodol*. 2007;69:659−77.

[50] Simon N, Friedman J, Hastie T, et al. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39:1−13.

[51] Tibshirani R. Univariate shrinkage in the Cox model for high dimensional data. *Stat Appl Genet Mol Biol*. 2009;8:3498−3528.

[52] Zhu J, Hastie T. Classification of expression arrays by penalized logistic regression. *Biostatistics*. 2004; 5(3): 427−443.

[53] O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics*. 1990; 46: 33−48.

[54] Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*. 2009: 361(10):947−57.

[55] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348−60.

[56] Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418−29.

[57] Zhang H, Lu W. Adaptive lasso for Cox's proportional hazard model. *Biometrika*. 2007;94:691−703.