COMPUTATIONAL MODELING FOR CENSORED TIME TO EVENT DATA USING DATA INTEGRATION IN BIOMEDICAL RESEARCH

by
ICKWON CHOI

Submitted in partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Thesis Advisors: Dr. Michael W. Kattan

Department of Electrical Engineering and Computer Science

CASE WESTERN RESERVE UNIVERSITY

August, 2011

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of

Ickwon Choi

candidate for the Ph.D. degree *.

(signed) Mehmet Koyuturk
(chair of the committee)

Michael W. Kattan

Andy Podgurski

Soumya Ray

(date) 06/08/2011

*We also certify that written approval has been obtained for any proprietary material contained therein.

Tables of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my research advisor Dr. Michael W. Kattan, Professor of Medicine, Epidemiology and Biostatistics, Cleveland Clinic Lerner College of Medicine of CWRU for his investment of time in this work and for his dedication and commitment to my Ph.D. study. Specially, I am very grateful to my academic advisor, Dr. Mehmet Koyuturk for his assistance and instruction to correct my dissertation. I would also like to thank Changhong Yu and Dr. Brian J. Wells of Quantitative Health Sciences, Cleveland Clinic for their help throughout the progress of this dissertation. Finally, I thank Dr. Andy Podkurski and Dr. Soumya Ray for serving on my dissertation committee and for their valuable mentorship.

List of Abbreviations

AIC	Akaike information criterion
BCD	breast cancer data
C-index	concordance index
CPE	concordance probability estimates
CV	cross validation
CVCI	cross validated *C*-index
CVPLL	cross-validated partial log-likelihood
CVPPLL	cross validated penalized partial log-likelihood
DCV	double cross validation
DD	difference in deviance
EPV	events per variable
FNSS	forward nested subset selection
FRR	feature relevance ranking
FSS	forward stepwise selection
GAMs	generalized additive models
IAUC	integrated area under the receiver operating curve
IBS	integrated Brier score
LDA	linear discriminant analysis
LOOCV	leave-one-out cross validation

LS-SVMs	least squares support vector machines
LRT	likelihood ratio test
MPLLE	maximum partial log-likelihood estimation
MPPLLE	maximum penalized partial log-likelihood estimation
NCI	Netherlands cancer institute
PCA	principal component analysis
PCR	principal components regression
PI	prognostic index
PLL	partial log-likelihood
PLS	partial least squares
PMLE	penalized maximum likelihood estimation
PPLL	penalized partial log-likelihood
ROC	receiver operating characteristic
STMC	maximum concordance index
SVD	singular value decomposition
UNOS	united networks for organ sharing

COMPUTATIONAL MODELING FOR CENSORED TIME TO EVENT DATA USING DATA INTEGRATION IN BIOMEDICAL RESEARCH

Abstract

by

ICKWON CHOI

Medical prognostic models are designed by clinicians to predict the future course or outcome of disease progression after diagnosis or treatment. The data, which are used when these clinical models are developed, are required to contain a high number of *events per variable* (EPV) for the resulting model to be reliable. If our objective is to optimize predictive performance by some criterion, we can often achieve a reduced model that has a little bias with low variance, but whose overall performance is improved. To accomplish this goal, we propose a new variable selection approach that combines Stepwise Tuning in the Maximum Concordance Index (STMC) and Forward Nested Subset Selection (FNSS) in two stages. In the first stage, the proposed variable selection is employed to identify the best subset of risk factors optimized with the concordance index using inner cross validation for optimism correction in the outer loop of cross validation, yielding potentially different final models for each of the folds. We then feed the intermediate results of the prior stage into another selection method in the second stage to resolve the overfitting problem and to select a final model from the variation of predictors in the selected models. Two case studies on relatively different sized survival data sets as well as a simulation study demonstrate that the proposed approach is able to

select an improved and reduced average model under a sufficient sample and event size compared to other selection methods such as stepwise selection using the likelihood ratio test, Akaike Information Criterion (AIC), and least absolute shrinkage and selection operator (lasso). Finally, we achieve improved final models in each dataset as compared full models according to most criteria. These results of the model selection models and the final models were analyzed in a systematic scheme through validation for independent performance evaluation.

For the second part of this dissertation, we build prognostic models that use clinicopathologic features and predict prognosis after a certain treatment. Most of the recent research efforts have focused on high dimensional genomic data with a small sample. Since clinically similar but molecularly heterogeneous tumors may produce different clinical outcomes, the combination of clinical and genomic information is crucial to improve the quality of prognostic prediction. However, there is lack of an integrating scheme into a clinico-genomic model due to the larger number of variables and small sample size, in particular, for a parsimonious model. We propose a methodology to build a reduced yet accurate integrative model using a hybrid approach based on the Cox regression model, which uses several dimension reduction techniques, L_2 penalized maximum likelihood estimation (PMLE), and resampling methods to tackle the problems above. The predictive accuracy of the modeling approach is assessed by several metrics via an independent and thorough scheme to compare competing methods. In breast cancer data studies for metastasis and mortality outcome, in a DLBCL data study, and in simulation studies, we demonstrate that the proposed methodology can improve prediction accuracy and build a final model with a hybrid signature that is parsimonious when integrating both types of variables. The selected clinical factors and genomic biomarkers are found to be highly relevant to the biological processes and can be considered as potential biomarkers for cancer prognosis and therapy. Furthermore, selected but unidentified genes are open to thorough investigation.

CHAPTER 1 Backgrounds

1.1 Prognostic prediction model

Prognostic prediction models are used in medical society for investigating clinical outcome in relation to patient and disease characteristics. Although many model building methodologies and computational tools exist in biomedical research, such final models built by using them do not always work well in practice. Thus it is needed to be validated before it is used as practical tools. In particular, uncritical application of modeling techniques can result in models that poorly or overly fit the dataset and inaccurately predict outcomes on new subjects. In this dissertation, we investigate two types of modeling design methodology for constructing a reliable final model using unbiased validation techniques. The two design methodology in primary problems are 1) to build a clinical prognostic model in a high *events per variable* (EPV) setting and 2) to build a clinicogenomic model using data integration in the high dimensionality and small sample size setting.

1.1.1 Clinical prognostic model

Medical prognostic models can be designed in a high *events per variable* (EPV) setting to predict the future course or outcome of disease progression after diagnosis or

treatment. Such models can provide individualized predictions about the characteristics of one single patient. However, there is considerable uncertainty within the statistical modeling community regarding how to develop an accurate prediction model for censored survival data [10]. Specifically, when it comes to variable selection, some advocate fitting the full model [22] in which predictors are pre-specified with external information from the literature, while variable selection methods remain popular [1,52]. Nonetheless, a full model may be large and complicated to be used as a statistical tool. There is little literature comparing these primary approaches with respect to the predictive accuracy in censored clinical data. Recently logistic regression models [52-55] have been studied for clinical models. If the goal is to optimize predictive accuracy for finding a set of reduced prognostic factors, a plausible alternative to the full model would be to fit the most accurate, possibly reduced, model. An argument can easily be made for a parsimonious model that is at least as accurate as the full model.

1.1.1.1 Problems and strategies

1.1.1.1.1 Estimation and selection bias in model selection

In general, the complexity of a model obtained by a procedure of model selection is expected to be less than that of the full model, and the variance of the estimated parameters should be lower. Nevertheless, recent studies emphasize the limitations of variable selection, such as bias in the estimates of parameters (*estimation bias*) and the

lack of stability in an iterative scheme of variable selection [53]. In a stable algorithm, the effect of computational error during the iteration is no worse than that of a small amount of input data error from a phenomenon in which two or more predictor variables in a model are highly correlated (multicollinearity) [13,26]. An unstable variable selection algorithm may enlarge initial perturbations after numerous iterations. Furthermore, in variable selection, multicollinearity between the omitted variables and the selected variables can cause *selection bias*. Dropping influential variables from the effective model results in underfitting to data with increased residuals and biased parameter estimates for selected variables (*omission bias*). Adding unimportant variables to the effective model induces overfitting and increases the variance of parameter estimates for correlated predictors [36]. In this study, we attempt to reduce the instability and increase the reliability of the selection algorithm using the resampling method of cross validation [48].

1.1.1.1.2 Model complexity

A large sample size is the need for a problem of fitting the full model with numerous and complicated predictors to obtain unbiased estimation in model fitting, and for the possibility of overfitting due to the high model complexity. The sample size problem due to the model complexity can be accounted for by the *curse of dimensionality* [3]. In fact, regression modeling with time to event data, which contain sequential time

information on event such as death of recurrence of a disease, is much more sensitive to the events per variable (EPV) [43] than the overall sample size. Some researchers carefully guide the ten EPV ratio to estimate bias and sometimes suggest using shrinkage of the coefficient estimates [54]. However, highly correlated features in this situation may produce high variance, even if there is no estimation bias according to the EPV. Hence, this guidance is crucial to model building at the development step.

1.1.1.1.3 Right censoring effect

The last challenging characteristic of clinical survival data, to tackle in variable selection, is right censoring. There are two types of censoring in classical survival models: (i) Type I: survival until the end of study but whose final event time is unknown; (ii) Type II: lost to follow-up after a certain time. Even though data are incomplete, they contain a certain amount of information to increase the sample size and thus improve performance of the model. However, with the presence of censoring, the behavior of the underlying mechanism produces unclear performance measurements of models and may lead to biased results in variable selection. In survival analysis, Cox regression models are commonly used, and one of the major advantages is the ability to utilize censored observations. We use the Cox proportional hazards model [10,40] in this dissertation. In order to consider the censoring effect in model assessment, several performance measures, some of which summarize a time dependency using integration [21,27] and are robust to

censoring [35], are introduced to quantify the prediction accuracy and the amount of prognostic information represented by the model. Among them, maximizing the concordance index (*C*-index) has some patterns to enhance other measures along with it and some merits (see Section 2.1). As a criterion for prediction accuracy, the *C*-index is a preferred choice in this study.

1.1.1.2 Model selection principle

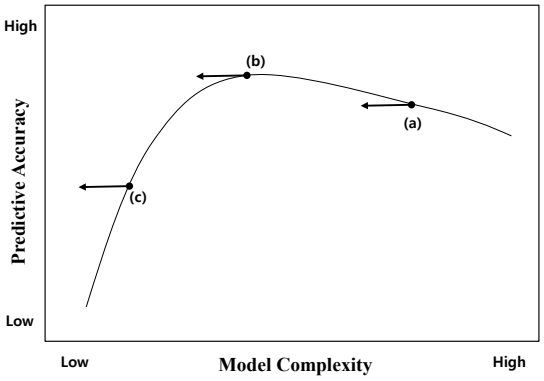


Figure 1.1 Types of initial full models in the optimization path to their final models

Figure. 1.1 illustrates the optimization path with the initial point of a full model in a variable selection procedure. The selection methods start from the full model, which is a type of single final model, and select the best model, optimized in some criterion. The starting full models can be categorized into 3 groups depending on the above challenges

with the data involving the event size, the model complexity, and the degrees of censorship: (a) model is more complex than the optimal model, (b) model is at optimum complexity, and (c) model is not adequately complex in Figure 1.1. The objective of model selection is to achieve the final model with optimal model complexity based on the prediction accuracy while tuning the tradeoff between bias and variance. In theory, the type (a) completes the course at (b), and in the types of (b) and (c), the full model is the final model, in which the difference is that in (c) the full model may suffer from a lack of time to event data, adequately significant predictors, or high rate of right censoring at the initial point.

1.1.1.3 Specific aim and proposed approach

The first aim of this dissertation is to propose a novel approach that builds a parsimonious model, in a high EPV setting, that is at least as accurate as the full model with respect to the *C*-index as an objective criterion. Herein, we propose a new approach to address these problems in two stages: (1) stepwise tuning in the maximum concordance index (STMC) as a variable selection process using inner cross validation for the optimism correction within each set of training folds of outer cross validation and (2) forward nested subset selection (FNSS) as overfitting control, which reduces uncertainty and variability in the predictors of chosen models resulting from STMC and builds a single final model. In the new approach, Cox proportional hazards regression

models with only main effect terms are used and fitted to two censored clinical data sets in the areas of renal transplantation and prostate cancer. For the comparative study of methods and models, we employ the same scheme as the first stage of our approach to compare our proposed method against the alternatives of the stepwise method that uses the likelihood ratio test and *Akaike information criterion* (AIC) criterion and the least absolute shrinkage and selection operator (LASSO) using an L_1 absolute value penalty that has two meritorious features of shrinkage and model selection [16,59]. Then, we compare the single final model of the FNSS result with the full model for final model assessment.

1.1.2 Clinicogenomic models for Data integration

1.1.2.1 Introduction

In clinical research, predictive models, such as nomograms [28,30], are developed based on clinical expertise or empirical results from the clinical literature, are validated externally, and are put into practical use for outcome prediction. High throughput gene expression profiles of primary cancer in the microarray technology have the potential to identify prognostic molecular markers associated with cancerous and metastatic phenotypes. Such findings can lead to translational research, the process of which translates those scientific discoveries into critical applications such as diagnostics, prognostics and treatments, and hence serves as a bridge between lab bench discoveries

and the patient bedside [17]. However, the deluge of both gene expression and clinical risk factor data is outpacing our ability to interpret and analyze using current models predictive for survival. It is increasingly apparent that a data integration scheme for model building that combines clinical and genomic variables is required to integrate data from heterogeneous sources, if maximal information is to be extracted and synergy created.

1.1.2.2 Literature review for integration scheme

The key challenge for integrative model building strategies in cancer prognosis is the high dimensionality and small sample size that characterizes microarray data, or the P >> N problem. Because the number of independent variables (P) far exceeds the number of individuals (N) in the training sample, model *overfitting* occurs, resulting in overwhelming *overoptimism*. Thus standard mutivariate Cox regression analysis cannot be directly applied to the data. Dimension reduction techniques of feature extraction, such as partial least squares [38] and supervised principal components [2], or feature selection (e.g., filter, wrapper, and embedded methods) [2,6], are required to reduce the number of features to a sufficient minimum. Although dimension reduction techniques identify candidate genes, those genes are highly correlated and penalization methods may be needed to adjust for overoptimism.

1.1.2.2.1 Classification issues

Few articles on integrative models have been published, and of these, most have employed classification based approaches, where patients are classified into high risk or low-risk groups. The researchers of [51] employed a logit transformation of the patient's 7-year disease-progression-free probability, calculated from an extensively validated postoperative nomogram, as the first single clinical variable of a combined model in the stepwise logistic regression procedure. Gene variables were included until optimal classification was achieved within a training set. This clinical study was a case-control design of prostate cancer recurrence in which controls were chosen based on a minimum of 5 years of follow-up without evidence of recurrent cancer, and none of the patients developed recurrence after 5 years. The authors [14] applied Bayesian networks to integrate two different data sources but the model was not tuned for classification. To predict breast cancer prognosis, researchers [12] proposed kernel methods using least squares support vector machines (LS-SVMs) to learn simultaneously from multiple data sources in three ways: early integration, intermediate integration, and late integration. The researchers [57] developed a new feature selection algorithm (I-RELIEF), in which the optimized objective function approximates the leave-one-out accuracy of a nearest-neighbor classifier, to identify a hybrid signature from clinical and microarray data and linear discriminant analysis (LDA) is used to estimate classification performance. These studies have showed that the combination of clinical and gene expression data can significantly improve prognostic specificity over either data type alone. However, a more

practical and appropriate strategy is needs to handle the heterogeneity present in clinical and gene expression data (see Section 1.1.2.3).

1.1.2.2.2 Regression issues

Clinico-genomic models for survival prediction based on the Cox proportional hazards regression have also been proposed, and evaluated with comparisons of various well-known prediction methods [6]. The researchers used three different gene expression data sets (breast cancer, diffuse large B-cell lymphoma, and neuroblastoma) to perform a systematic comparison of the performance of prediction models using clinical covariates only, genomic data only or a combination of the two. They used the breast cancer dataset (295 women) reported by [64], in which they reduced the initial set of 24,885 genes to 4919 by employing the Rosetta error model as a first screening [65] and applied a global test for survival data developed by [18], which simultaneously tests the significance of all genes in a prognostic regression model, and obtained a highly significant result for the outcome of mortality (P–value < 0.00001). When building prediction models, both types of covariates are used simultaneously to infer the parameters, but dimension reduction was applied only to high dimensional genomic variables. They concluded that their clinico-genomic model using ridge regression outperformed six other methods in all three data sets. The researchers [34] performed gene ranking with a single gene and a multivariate set of clinical prognostic factors that is an adjustment, in the Cox regression

model and identified a set of significant genes based on the predictive accuracy in the multivariate Cox model of the standard prognostic factors for adjustment and a compound variable of genes for prognostic index instead of the criteria to control false discovery rates in multiple testing.

1.1.2.3 Problems and a proposed methodology

Usually, as cancer data analyzed in survival studies are collected with a censored time to event manner, if 1) two discrete classes, relapse or death based on a primary tumor are used as outcome response, 2) a compound variable such as a prognostic index rather than clinical variables, which is utilized as a single effect [32] or 3) gene expression profiles are transformed into categorical variables suitable for classification methods (e.g., as up or down-regulated expression), then underlying information might be lost. Therefore, in order to make the best use of data, we focus on the widely used Cox proportional hazards model [11] as a regression problem in survival analysis.

For the second part, we propose a methodology to extend our approach to clinical model building [9] to an integrative model using clinic-genomic information through double cross validation, considering the problems that integrative methods might have in dealing with the P>>N problem as well as the issues associated with Cox regression models such as right-censoring, event per variable [42], and co-linearity between clinical factors or molecular biomarkers. We propose preliminary univariate screening using a

permutation test to reduce an initial immense set of gene expression profiles. During the validation procedure, the Stepwise Tuning in the Maximum Concordance Index method (STMC) [9] is employed. If the model size is greater than the sample size, the dimensionality is reduced by QR decomposition, and in the Forward Nested Subset Selection (FNSS), L_2 penalized maximum likelihood estimation (PMLE) is used to adjust for overoptimism and only genomic variables are shrunk by the PMLE. This approach identifies a parsimonious set of relevant clinical prognostic factors and genomic signatures. To address the variation in censoring effect, competing methods are compared using various performance metrics. To ascertain the reliability of our experimental results, the integrative prognostic model is internally validated using double cross validation (DCV) to obtain generalization performance in a breast cancer data set [8] and final models are rigorously assessed by several accuracy measures through leave-one-out cross-validation (LOOCV). For further demonstration, a diffuse large B-cell lymphoma (DLBCL) data study and simulation studies were also performed. First, we define standard Cox proportional hazards regression using censored time to event data widely used in survival studies.

1.2 Censored time to event data and standard Cox regression

Censored survival data is defined as $\boldsymbol{z}_i = (t_i, \delta_i, \boldsymbol{x}_i)$ for n independent individuals, $i \in \{1, \dots, n\}$. The observed time, t_i is given by $t_i = \min(T_i, C_i)$, where T_i is the time

of event and C_i is the time of censoring. The event indicator variable, δ_i is equal to 1 if an event occurred at the observed time ($T_i \leq C_i$), or a value of 0 if t_i is censored, and $\boldsymbol{x}_i$ is a p-vector of covariates, $[x_{i1}, x_{i2}, \ldots, x_{ip}]^T$, which constitutes a row of the design matrix $\boldsymbol{X}$. The Cox proportional hazards model is defined as

$$h(t|\boldsymbol{x}_i) = h_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}), \tag{1-1}$$

where $h_0(t)$ is a baseline hazard function and left unspecified with $h_0(t) \geq 0$. For an estimate of the baseline hazard $h_0(t)$, the Breslow estimator is commonly used and given by

$$\hat{h}_0(t_i) = 1/\sum_{t_j \geq t_i} \exp(\boldsymbol{x}_j^T\boldsymbol{\beta}). \tag{1-2}$$

The vector of regression coefficients, $\boldsymbol{\beta}$, is estimated by maximizing the partial log-likelihood (*PLL*)

$$PLL_{full}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{x}_i^T\boldsymbol{\beta} - \log\left(\sum_{t_j \geq t_i} \exp(\boldsymbol{x}_j^T\boldsymbol{\beta})\right)\right]. \tag{1-3}$$

Our proposed methodology is based on the Cox model and will be designed to overcome the problems we addressed in this chapter.

CHAPTER 2 An empirical approach through validation for Clinical Models in a High Events per Variable Setting

Medical prognostic models can be designed in a high *events per variable* (EPV) setting by clinicians to predict the future course or outcome of disease progression after diagnosis or treatment. In this situation, measurement of predictive accuracy can be difficult for survival time to event data and is very important in the model design procedure. We first examine several performance measures to obtain unbiasedly validated predictive accuracy in a proposed comparative scheme.

2.1 Performance measures of censored time to event data for model and method assessment

There are many metrics used to measure a computational prognostic model's prediction accuracy. These are principally categorized into (1) *Discrimination*, which measures how well the prediction model can discriminate between cases with events and those without events, includes the time dependent receiver operating characteristic (ROC) curve [25], the concordance index (or *C*-statistic) [22,23], and the CPE (Concordance Probability Estimates) [18]; (2) *Calibration*, which quantifies how close a predicted estimate is to the real probability, includes the calibration slope and curve; and lastly (3)

Overall score of the measures such as explained variation (R^2 type statistics and the Brier score) [27]. Some relatively new performance measures for reclassification and clinical usefulness are also discussed as well as the above in [56] and they stress that a well-discriminating model may be most relevant for research purposes, suggesting that reporting discrimination and calibration is important for a prediction model.

Although the partial log likelihood (*PLL*) is used for predictive inference and modeling, it can also be used in the difference in deviance (DD) between a fitted model and the null model, given by $-2(PLL(\boldsymbol{\beta}) - PLL(\mathbf{0}))$ and can be used as a prediction error for evaluating the performance on new data as well as for the selection of complexity on training data.

As for explained variation, the variants of R^2 statistic on censored data can be defined in several ways [27] and are very sensitive to the rate of censoring, and it is tricky to determine which type of measure is proper for comparisons due to the uncertainty of a censoring mechanism.

Our concern in this study is to gauge the prediction ability of a model, whose optimism is corrected and whose estimation is unbiased, using the performance measures. However, numerous existing predictive measures have advantages and disadvantages for survival analysis and we would not insist that there is one that is superior to the others and the development of new measures is also an active area of research. Hence, we use several metrics, defined in the following sections, simultaneously in order to compare the performances of the different model selection methods, whose results may directly be

amenable to produce the empirical prediction performance for new patients [29].

2.1.1 Integrated area under the receiver operating curve

The ROC curve for the discriminative ability is a standard technique to assess the trade-off between sensitivity and 1-specificity in a binary classification rule [43]. The ROC curve is a plot of sensitivity and 1-specificity for all of the possible cutoff values, c of a continuous variable, which is the risk score R, such as the prognostic index (PI), which is a linear predictor, in survival analysis. The time-dependent ROC curves were proposed to assess the predictive accuracy of survival models [25], defined as

$$Sensitivity(c,t) = \Pr\{R > c | D(t) = 1\}, \qquad (2\text{-}1)$$
$$Specificity(c,t) = \Pr\{R \leq c | D(t) = 0\}.$$

Here, $D_i(t) = 1$ if $T_i \leq t$ and $D_i(t) = 0$ if $T_i > t$ and it represents the event status of individual i at time t. The corresponding time-dependent ROC curve and the time-dependent area under the ROC curve can be defined for time t as ROC(t) and AUC(t), respectively. The AUC(t) can be summarized by the integrated IAUC, given by the area under ROC(t) over event time. As for the AUC, an IAUC=1 indicates perfect prediction accuracy and IAUC =0.5 is as good as a random guess over time.

2.1.2 Concordance index

In survival analysis, one of the most popular performance measures for assessing models is the concordance index, which is similar to the Wilcoxon-Mann-Whitney statistic in bi-partite ranking problems [33]. The concordance index [1] is defined, for the second measure of the discriminative ability, as

$$C\text{-index}=\frac{\sum_{i,j\epsilon\Omega}\mathbf{1}\{PI_i<PI_j\}}{|\Omega|}. \quad (2\text{-}2)$$

Here, PI_i and PI_j is a linear combination of clinical variables and their estimated coefficients (a linear predictor) of patient i and j. $\mathbf{1}\{\}$ is an indicator function that is equal to 1 if the argument is true, 0 if false, and Ω is the set of all pairs of patients (i, j) that satisfy one of the following: (i) the patients i and j experienced recurrence and the recurrence time t_i is shorter than t_j, or (ii) only patient i experienced recurrence and t_i is shorter than the follow-up time t_j. Tied pairs contribute 1/2 weight to the numerator and denominator. The C-index estimates the probability that given two randomly selected patients, the patient who has the event first also had a higher probability of the event. The experienced recurrence time of an individual in (2-2) can be replaced with the prognostic index. The C-index is a metric to compute and measure discriminative ability utilizing a complete dataset.

Although the *C*-index is unable to represent evolutionary performance over time, it is a generalization of AUC(*t*) [21]. Also, the researchers in [44] demonstrate that a method maximizing the *PLL* ends up approximately maximizing the concordance index. We have thus chosen the concordance index as the primary objective criterion in our proposed approach due to its popularity, interpretability, simplicity, and robustness, though several measures in this section will be utilized to compare methods and models on account of the censoring variation.

2.1.3 Calibration slope and curve

Calibration is performed by using the calibration slope and calibration curve. The calibration slope β for survival data can be computed by performing a Cox regression with the PI (prognostic index) for a new data set, as a single continuous variable in the Cox proportional hazards model as follows.

$$h(t|PI) = h_0(t)\exp(\beta \cdot PI). \tag{2-3}$$

Here, the prognostic index is a linear combination of the regression coefficients estimated in a training sample and the values of risk factors in the test data. If the calibration slope is unity, the regression model is perfectly calibrated. Otherwise, the

regression coefficients that are estimated in the training sample reflect underestimation or overestimation. For the validity of the whole model, however, we need to check the correctness of the baseline survival function as well [63] and the optimism corrected slope can be considered as a shrinkage factor that takes overfitting into account [23]. Calibration is also visually inspected by a calibration curve which is a plot of groups with their equal sample sizes and displays the accuracy between average predicted probabilities vs. Kaplan-Meier estimates of actual outcomes.

For performance evaluation and model validation, two main concepts of discrimination and calibration can be combined for a data analysis. These can provide a complementary interpretation for comparative analysis, as the overall score is suffering from a censoring mechanism.

2.1.4 Integrated Brier score (IBS)

For the inaccuracy of individual predictions, the censored brier score (CBS) is calculated based on the sum of squared differences between predicted and observed survival with censorship [19]. CBS can be computed empirically as a function of time *t* for *n* patients of multiple covariate $\boldsymbol{x}$ with a censoring variable δ_i and a time to event variable T_i as follows.

$$CBS(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(0-\hat{\pi}(t|\boldsymbol{x}_i)\right)^2\mathbf{1}(T_i \le t, \delta_i = 1)\left(\frac{1}{\hat{G}(T_i)}\right)+\left(1-\hat{\pi}(t|\boldsymbol{x}_i)\right)^2\mathbf{1}(T_i > t)\left(\frac{1}{\hat{G}(t)}\right)\right\}. \quad (2\text{-}4)$$

Here, $\hat{\pi}(t|\boldsymbol{x}_i)$ is an estimated recurrence free probability for a patient i, and $\hat{G}(t)$ is a probability of being censoring and is calculated by the Kaplan Meier estimate on $(T_i, 1-\delta_i)$. The variable T_i has the time value of an event recurrence if the event status of a patient i, δ_i is 1, or a censored time if δ_i is 0. Note that the Brier score is 0 in the perfect prediction and 0.25 when the trivial prediction of $\hat{\pi}(t)$=0.5 is made for all patients. The integrated Brier score (IBS) is a summary of the prediction error over event time by integrating the formula (2-4).

2.2 Risk prediction methods

For the comparison purpose with our proposed method, first we introduce the stepwise selection method starting from the full model using two different criteria, the likelihood ratio test and *Akaike information criterion* (AIC) in Cox models, and we define the LASSO method that perform variable selection while penalizing coefficient parameters in the following section.

2.2.1 Stepwise variable selection and *P*-value

The standard stepwise variable selection method uses a p-value to

determine a variable to be eliminated from or be inserted into an accurate model. The p-value is based on a statistical hypothesis test in which a result is statistically significant if it is unlikely to have occurred by chance alone, according to a pre-specified threshold probability, the significant level.

2.2.1.1 Hypothesis test

In order to test the null hypothesis $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$, where $\boldsymbol{\beta}^{(0)}$ are the coefficients of the null model or nested smaller model with p_0 parameters in contrast to the bigger model with the p_1 parameters of $\boldsymbol{\beta}$, Wald test or the likelihood ratio test can be used to derive the significance of a variable for variable selection. The probability distribution of both test statistics is approximated by a chi-square distribution with $(p_1 - p_0)$ degrees of freedom.

2.2.1.2 Wald test

The Wald test statistic is written as $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})^T \hat{\boldsymbol{I}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})$, where $\hat{\boldsymbol{I}} = \boldsymbol{I}(\widehat{\boldsymbol{\beta}})$ is the estimated information matrix, which is the second derivative of the log partial likelihood (*PLL*) with respect to $\boldsymbol{\beta}$. This reduces to the z-statistic $\hat{\beta}/se(\hat{\beta})$, where $se(\hat{\beta})$ is a standard error of $\hat{\beta}$, in a single variable [16,63].

2.2.1.3. Likelihood ratio test (LRT)

The likelihood test statistic is defined as $2(PLL(\widehat{\boldsymbol{\beta}}) - PLL(\boldsymbol{\beta}^{(0)})$, where $PLL(\widehat{\boldsymbol{\beta}})$ and $PLL(\boldsymbol{\beta}^{(0)})$ are the log partial likelihood of the bigger model and the nested smaller model respectively. The likelihood test is generally more stable than the Wald test, and therefore is used for the stepwise selection.

2.2.1.4 Stepwise selection in Cox model

The stepwise variable selection is a hybrid method in either direction of two classic methods: (1) the forward selection method and (2) the backward elimination method, and uses the significance of a variable as a criterion for selection. In Cox proportional hazards regression models, the likelihood ratio test is used to compute p-values. In the forward selection, the algorithm begins with the null model and adds the predictor with the smallest p-value. This is repeated until no variable upon entry into the model has a p-value less than a significance level that is a parameter to be determined in advance. If variables entered are no longer significant they may be dropped off while candidate variables are added in the forward approach. This step is repeated until the final model has no variables with p-values greater than or equal to the significance level. For the backwards elimination, we start with the full model and remove the variable with the

largest p-value if it is larger than the significance level. We repeat this process until no variable in the model has a p-value greater than or equal to a significance level. The hybrid stepwise gives a second chance to each dropped variable except the most recently dropped one. If this value is more significant than that of the dropped one and less than the significance level, the predictor is reintroduced in the current model of the process. The hybrid stepwise is used in this dissertation and we refer to it as stepwise selection. As a threshold significance level for selection, the conventional quantity 0.05 is used.

2.2.2 AIC for variable selection

The AIC for variable selection is defined by

$$AIC = -2 \cdot PLL(\boldsymbol{\beta}) + 2 \cdot d, \tag{2-5}$$

where d is the effective number of parameters and given by the number of parameters in a model. We simply choose the model giving the smallest AIC over the subsets of models considered in each search space started from the full model. The AIC estimates prediction errors in an analytical and intrinsic way that the optimism is estimated directly from a training set and then this is added to the training error. So the optimism correction using cross validation or bootstrapping is not required in such a criterion. Although future inputs are not likely to be identical to training sets, this kind of error can be used for the

effective model selection due to its relative nature. It turns out that the significance level 0.157 in the backward elimination method usually chooses the variables that are selected by minimizing the AIC in all subset procedure when all variables have 1 degree of freedom [1]. This criterion is used in the stepwise selection method instead of a *P*-value.

2.2.3 Lasso

The last baseline method we choose for comparison is the L_1 penalized estimation method, LASSO [16,59], that shrinks the estimates of the coefficients of a Cox model towards zero by imposing a penalty on their absolute values. It has a built-in feature selection procedure while penalizing the parameters unlike L_2 penalized Cox regression with a quadratic penalty (ridge regression) [5,6] that allows all coefficients to be non-zero and may yield complex models. The objective of this shrinkage is to prevent overfitting occurring by collinearity of the covariates. Thus we fit the parameters $\boldsymbol{\beta}$ of clinical variables $\boldsymbol{x}_i$ for patient i by maximizing L_1 penalized partial log-likelihood (PPLL) defined over the entire data with an absolute value (lasso) penalty λ on $\boldsymbol{\beta}$ as follows.

$$PPLL_{\lambda}(\boldsymbol{\beta}) = PLL_{full}(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1 \tag{2-6}$$
$$= \sum_{i=1}^{n} \delta_i \left[(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \log\left(\sum\nolimits_{t_j \geq t_i} \exp\left(\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta}\right) \right) \right] - \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda > 0$ and $\|\cdot\|_1$ stands for the L_1 norm. The zero value of λ means no shrinkage and the infinity value indicates infinite shrinkage. In our study, we used the R package *penalized* to apply the lasso implementation and used likelihood cross-validation for optimizing the tuning parameter [16].

2.3 Proposed approach

2.3.1 Comparative Scheme for the unbiased assessment of methods and models

The framework of model building generally consists of (1) model selection for the final model to find a final set of predictors or determine tuning parameters for model complexity, (2) validation and assessment using internal validation for the final model and model selection methods, and (3) the final model building for the practical use; including learning the parameter coefficients of the predictors that are found in the previous steps and its application of the final model to the external data sets (external validation).

For building a final model using model selection, the common procedure is to (1) make best use of the whole data set for finding model complexity parameters or for identifying a final set of variables, instead of using a test data held out for validation, and (2) proceed to estimate the coefficients of predictors, also using the data, for the single

final model.

As for the unbiased assessment of the final model, we need a final set of predictors, the resultant tuning parameters of the model complexity in learning methods, and resampling techniques, such as data splitting, cross validation (CV), or bootstrapping. This assessment scheme should be differentiated from that of model selection methods for comparisons we used in the proposed methodology as below.

The relative performance of a model within a variable selection method may be subject to the variability of the training data on account of the EPV, selection bias, and right censoring in survival data. Thus, we need the unbiased estimate of the true performance of a variable selection method, and it can be achieved, using CV, by the fact that all the aspects of the model development such as model selection and parameter tuning should take place in the training sets within the CV [66]. Although the Leave-one-out cross-validation (LOOCV) and bootstrapping, in general, perform well regarding bias, *n*-fold CV may be preferable to them due to the lower computational cost [37]. For this reason, we employ the CV as an outer loop for the assessment of variable selection methods and a nested CV of training folds within the outer CV for the optimism correction. In order to make the best use of the training data in variable selection, we randomly permute the data for the repeated resampling and obtain the replicates of CV. Each of the *n*-final models of *n*-fold CV after model selection is tested on the fold set left out for the independent evaluation and they are averaged for the assessment of each method. The procedure of this scheme is displayed in Table 2.1. For the best optimized

final model, note that it is crucial to hold the ten EPV guideline in the developing step of model selection.

The stepwise methods with the likelihood ratio test and the AIC are applied to this scheme without the resampling setting of inner validation to compare with the proposed method.

Table 2.1 The procedure for evaluating the performance of variable selection methods

1. For each variable selection method:
   i) For each training data set obtained from n_1-fold outer cross validation:
      - Find the optimal set of predictors using k replicates of n_2-fold nested cross validation using the randomly permuted training data set.
      - Given the set, estimate the vector of regression coefficients $\boldsymbol{\beta}$ on the entire training data set.
      - Compute the values of several performance criteria (see section 3.2) on the test data set held out from outer cross validation.
   ii) average the performance measures of n_1 final models built from i).
2. Compare the different variable selection methods in terms of the criteria values.

2.3.2 Final model building through validation

Our approach for variable selection consists of two stages. First, Stepwise Tuning in the Maximum *C*-index (STMC) begins from a full model using the backward elimination. After a round of elimination, it reanalyzes the discarded variables one by one

and allows one more chance to be included in the current model. Using the n_1-fold outer CV, we obtain a set of n_1 final models with the different sets of predictor variables fitted to a training subsample and optimized, using the k- replicates of n_2–fold inner CV, for the maximum C-index. This might approximately represent the proportion of predictors in the final model and is used for the interpretation of their relative importance. As an overfitting control, numerous methods use a regularization scheme (e.g. weight decay in Neural Networks), early stopping during repetition, or a Bayesian approach. Instead of using early stopping in the first stage, we achieve this effect in the second stage of forward nested subset selection (FNSS) using variable ranking from the results of the distribution.

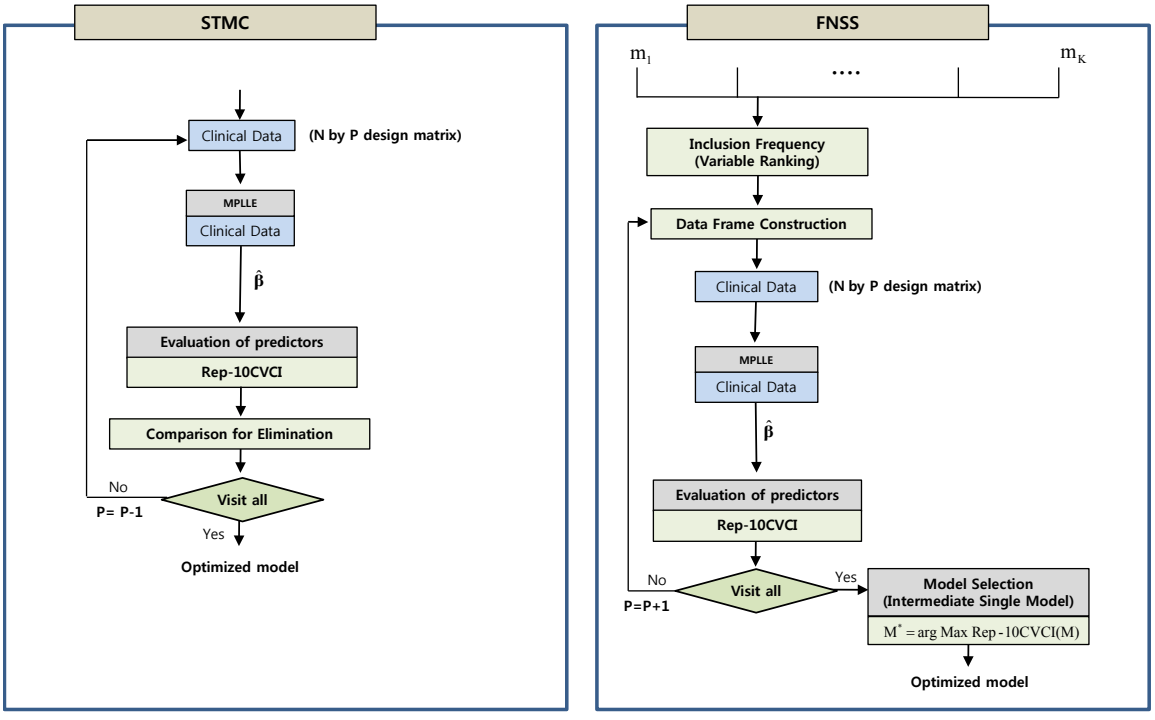


Figure 2.1 Flow diagram for the STMC and FNSS method.

Figure 2.1 shows a flow diagram for the STMC and FNSS method and will be described in the following sections.

2.3.2.1 First stage: STMC (Stepwise Tuning in Maximum *C*-index))

The STMC method we propose is in the class of wrapper methods [4,31]. The *C*-index is used for the performance evaluation of a subset of predictors in the goodness of fit. This kind of method might produce correlated variables in a group of predictors of a chosen model, but have the optimal performance. The iterative process of variable selection can be viewed as performing the optimization search in the model space. The search is performed over finite models in a given model class for model selection. Given p predictors, there are 2^p possible variable subsets for the entire search. It is usually very expensive to compare all combinations for a large p and so typically some heuristic search procedure is used to find a locally optimal good feature subset. We design the STMC method based on the backward elimination scheme and give one more chance for variables to reenter the model in the forward direction. This process is repeated until all variables are visited.

Table 2.2 The algorithm of STMC

Algorithm STMC

- Input : F, the set of variables in the full model
- Output : max, the final set of variables in STMC

1. Initialize $F = \{1,\ldots,p\}$, where p is the number of predictor and *visit* = $\{\varnothing\}$, and let max = F be the best subset.
2. Repeat
   {
   // a. Drop Step
   for i = 1 to p
   {
   - if ($i \in F$), $F_i = F - \{i\}$
   - CVCI(F_i, k, N) // Estimate k-replicated N-fold cross validated C-index to evaluate only the predictors of F_i in Cox Model.
   }

   // Compare the best feature subset found in Drop Step with *max*, and set *max* to be the subset with the greater C-index.
   - if (MAX(CVCI(F_i, k, N))> CVCI(*max*,k, N)) *max* = MAX_{Fi}(CVCI(F_i, k, N))
   - *visit* = *visit* $\cup$ MAX_i(CVCI(F_i, k, N))
   - $F = F$ - MAX_i(CVCI(F_i, k, N))

   // b. Add Step
   for j = 1 to length(*visit*)
   {
   - if ($j \in$ *visit*), $F_j = F + \{j\}$
   - CVCI(F_j, k, N) // Estimate k-replicated N-fold cross validated C-index to evaluate only the predictors of F_j in Cox Model.
   }

   // Compare the best feature subset found in Add Step with *max,* and set *max* to be the subset with the greater C-index.
   if (MAX_j(CVCI(F_j, k, N))>CVCI(*max*, k, N))
   {
   - *max* = MAX_{Fj}(CVCI(F_j, k, N))
   - *visit* = *visit* - MAX_j(CVCI(F_j, k, N))
   - $F = F \cup MAX_j$(CVCI(F_j, k, N))
   }

   // c. Stopping Rule Check
   - if (F=={$\varnothing$}) break the Repeat iteration
   }
3. Select and return the best feature subset, max which is evaluated during the search space

* MAX(CVCI(F_i, k)): a maximum value of CVCI(F_i, k)
* MAX_i(CVCI(F_i, k)): i with a maximum value of CVCI(F_i, k)

The pseudo code of this procedure is shown in the STMC algorithm of Table 2.2. Initially, a full model is assigned as a current best model with the maximum C-index, and

a set of visited variables is initialized to be empty. The repeat loop in the STMC algorithm is comprised of 1) Drop step, 2) Add Step, 3) Comparison for choosing an intermediate best model, and 4) Stopping rule check (when all variables are searched, break the iteration). The drop step (backward direction) tests each predictor by comparing the current model with a potential model whose size is one smaller than the current model, and eliminate the most irrelevant predictor producing the smallest *C*-index in the current best feature set when excluded from the current model. If there are no variables to win over the current feature, no changes happen in the best model. In the add step (forward direction), every element in the visit set of discarded predictors is given a possibility to be reintroduced in the current best model except for the element extracted from the previous drop step. Between both models from the drop and add step, the set of predictors with the greater *C*-index is chosen. The repetition stops when the procedure considers all predictors in the pool of feature variables.

The n_1-fold outer CV is used for the investigation of variable selection having the uncertainty of different predictors and sizes. This should be distinguished from the k replicates of randomly permuted training data sets and their internal n_2-fold CV, with which the generalization *C*-index of each potential model is evaluated by the sample re-use of *k*-replicates of the internal CV for the purpose of the overoptimism correction. The STMC method builds n_1 intermediate models and yields the distribution of predictors for a final model. The inclusion frequency is computed by the proportion of each variable in those models and reflects the significance of variables in the distribution. The results of

STMC are connected to the filter type approach of FNSS at the next stage. Our approach is very complex in running time because it performs internal validation while building intermediate models. The time complexity of the STMC algorithm is $O(k \cdot n_2 \cdot N \cdot P^2)$.

2.3.2.2 Second stage: FNSS (Forward Nested Subset Selection)

The proposed Forward Nested Subset Selection (FNSS) algorithm is a filter type method [20], and is designed for controlling overfitting caused from model selection and for identifying a single final model. The ranking criterion is defined for individual variables by the inclusion frequency obtained in the previous stage. High score variables are regarded as valuable, and they are sorted in the decreasing order of the inclusion frequency. After variable ranking, the FNSS builds models with increasing numbers of predictors while incorporating a variable one by one from the null model and evaluates each constructed model through the 10-fold cross validation based on the *C*-index, the IAUC, the calibration slope, the calibration curve, and the IBS. As our approach chooses the *C*-index as an objective metric, we select a set of variables with the maximum *C*-index as the final model. The time complexity of the STMC algorithm is $O(n_2 \cdot N \cdot P)$.

The R software version 2.8.1 [45] with the Design and survcomp packages were used to perform all analyses and the proposed approach is implemented with R package for free use (https://vorlon.case.edu/~ixc27/).

2.4. Results of two case studies

We first describe two datasets of prostate cancer and renal transplantation and apply our methodology to them to compare predictive performance with other methods.

2.4.1 Datasets

2.4.1.1 Prostate cancer data

We procured data from a study that created a postoperative nomogram for predicting the risk of prostate cancer recurrence [30] following institutional Review Board waivers (Cleveland Clinic IRB number: 4270). The cohort consists of a total of 1123 patients (with 167 biochemical recurrences) with clinically localized prostate cancer treated with open radical retropubic prostatectomy between 1987 and 2003. The seven predictors in the full model include the following categorical variables: (1) svi (seminal vesicle involvement), (2) sm (surgical margins), (3) lni (lymph node involvement) and (4) ece (extra-capsular extension), and the continuous variables: (5) psa (prostate specific antigen), (6) experience (surgery experience), and (7) pgx (postoperative Gleason sum) which is treated as an ordinal type variable. In [9], the full model is prespecified based on medical literature reviews and clinical knowledge of investigators and surgeons prior to an analysis of the data. For the further detail of the description of the data, see [9]. Two

missing values in psa are imputed using the R MICE package in the study and other variables are complete. Patients who are lost to follow-up or died from causes other than prostate cancer are right-censored. Table 2.3 shows the statistical description of the prostate cancer recurrence data in our study, and the estimated coefficients and statistical significance of the predictors in a multivariable Cox proportional hazards regression fitted to the entire data set for the full model and the final model built from the proposed method, which predict the 10-year probability of freedom from cancer recurrence defined as a PSA level > 0.4 ng/mL and rising, or a secondary treatment for a detectable and rising PSA less than or equal to 0.4 ng/mL.

Table 2.3 Description of prostate cancer data (1123 patients), and estimated coefficients and statistical significance of predictors in a multivariable Cox Proportional hazards model fitted to the entire data for the full model and the final model built by the proposed method. $\hat{\beta}_{full}$, estimated log-relative risk (full model, 7 predictors); P_{full}, P-values of full model; $\hat{\beta}_M$, estimated log-relative risk (model *M*: STMC+FNSS, 5 predictors); P_M, P-values of *M*.

Predictor	**No.(per cent)**	$\hat{\beta}_{full}$	P_{full}	$\hat{\beta}_M$	P_M
Pathology Gleason Sum		0.95	<0.00001	0.92	<0.00001
4-6	449(40)				
7	621(55)				
8-10	53(5)				
ExtraCapsular Extenstion(yes)	389(35)	0.92	<0.00001	0.95	<0.00001
Surgical Margin(yes)	297(26)	0.63	0.00019	0.65	<0.00011
Seminal Vesicle Involvement(yes)	89(8)	0.29	0.19	0.42	<0.048
Lymph Node Status(Positive)	23(2)	0.56	0.048		
PSA*(ng/mL)	(0.5) 7.6 (94.5)	0.02	0.033	0.02	0.0066
Surgery Experience*	(0) 679.2 (1336)	-0.00	0.67		

* continuous variable: (Min) Mean (Max)

2.4.1.2 Renal transplantation data

Renal transplantation data was obtained from the UNOS (United Networks for Organ Sharing) Registry for chronic kidney disease from 2000 to 2003 [60] and appendix A describes the original renal transplantation data of 20085 patients and 67 variables. The cohort includes 20085 living donor renal transplant cases with 2,300 documented graft failures. This data is used to form pre-transplant and post-transplant nomograms that predict 5-year graft failure in [60], in which all patients received kidney transplant as a primary treatment for renal failure and are then followed for signs of the transplant failure. The outcome of transplant failure is defined as a recurrence of kidney disease within 5 years of transplant. In the study, the predictor variables for the full models are chosen by clinicians based on their theoretical association with graft failure in the clinical literature.

Our study is based on the post-operative nomogram [60] and we use the data with 22 variables selected based on the full model [60] that is specified by clinicians from the 67 original predictor variables, which include some measurements before and after the time of the renal transplant. The predictors consists of demographical information of donors and recipients: age, gender, race (black, white, and others); pathological information: bmi (Body Mass Index), Donor Serum Creatinine Pre-transplant (SCr), Donor Procedure, Nephrectomy Type, HLA (Human Lymphocyte Antigen) mismatch level, Dialysis in the first week, any treatment for rejection within first 6 months, eGFR (Estimated Glomerula Filtration Rate)-MDRD (Modification of Diet in Renal Disease) after 6 months of transplant, Adjuvant chemotherapy on the use of immunosuppressants:

(Azathioprine, Rapamicine (Sirolimus), Mycophenolate Mofetil), IL2 Receptor Antibodies, Calcineurin Inhibitor without fk506, and Induction with Depleting Antibodies. The race variables of donors and recipients have a categorical type and are processed using dummy variables and HLA Mismatch is treated as an ordinal type. No missing values are identified.

The final predictive model predicts the 5-year graft survival probability after living donor kidney transplant. Table 2.4 further describes the above predictors and multivariate analysis used for the prespecified full model of the multivariable Cox model and the final model resulting from the proposed method.

Table 2.4 Description of renal transplant data (20085 patients), and estimated coefficients and statistical significance of predictors in a multivariable Cox Proportional hazards model fitted to the entire data for the full model and the final model built by the proposed method. $\hat{\beta}_{full}$, estimated log-relative risk (full model, 22 predictors); P_{full}, P-values of full model; $\hat{\beta}_M$, estimated log-relative risk (model *M*: STMC+FNSS, 7 predictors); P_M, P-values of *M*.

Characteristic	Variables	Description	Mean(SD) or No.(%)	$\hat{\beta}_{full}$	P_{full}	$\hat{\beta}_M$	P_M
Recipient	Age	Recipient Age(yrs)	46(14)	0.003	0.129		
	Gender	Recipient Gender(Female)	8,320(41)	-0.05	0.242		
	race	Recipient Race					
	Black		2,992(15)	0.24	0.09	0.424	<0.0001
	White		13,525(67)	-0.03	0.756		
	Others		3,568(18)				
	bmi	Recipient Body Mass Index(kg/m^2)	26.7(5.4)	-0.002	0.525	0.0001	0.034
Donor	AGE_DON	Donor's Age(yrs)	40(10.8)	-0.005	0.02	-0.004	0.978
	GENDER_DON	Donor's Gender(Female)	11,806(59)	0.003	0.941		
	drace	Donor Race					
	Black		2,767(14)	0.174	0.233		
	White		13,876(69)	0.022	0.834		
	Others		3,442(17)				
	d_bmi	Donor's BMI(kg/m^2)	26.9(4.7)	0.013	0.0029		
	d_creat	Donor Serum Creatinine(SCr) Pre-Tx(mg/dl)	0.9(0.5)	-0.081	0.148		
	d_procec	Donor Procedure:Nephrectomy Type		-0.017			
	Laparoscopi		13,057(65)		0.148		
	Open		7,028(35)				
Recipient/Donor	HLAMIS	HLA Mismatch Level		0.021	0.106		
	0		2,148(11)				
	1		1,262(6.3)				
	2		3,738(19)				
	3		5,739(29)				
	4		2,500(13)				
	5		2,956(15)				
	6		1,538(7.7)				
Adjuvant Chemotherapy	im_deple	Induction with Depleting Antibodies(yes)	3,731(19)	0.013	0.817		
	im_il2	Induction with IL2 Receptor Antibodies(yes)	7,604(38)	-0.013	0.77		
	im_aza	Azathioprine Maintenance(yes)	723(4)	-0.142	0.223		
	im_myco	Mycophenolate Mofetil Maintenance(yes)	15415(77)	-0.318	<0.0001		
	im_rapa	Rapamycin(Sirolimus) Maintenance(yes)	2960(15)	-0.324	<0.0001		
	im_calci	Calcineurin Inhibitor with fk506(yes)	18729(93)	-0.934	<0.0001	-0.903	~0.00000
Recipient Posttranplant	dial_1wk	Dialysis in the First Week(Yes)	957(5)	1.423	<0.0001	1.403	~0.00000
	trt_rej6	Any treated for Rejection within 1st 6mths(Yes)	1,861(13)	0.425	<0.0001	0.423	~0.00000
	gfr_po6	eGFR(MDRD) in 6 Mths(ml/min/1.73m^2)	56.5(18)	-0.025	<0.0001	-0.025	~0.00000

2.4.2 Experimental results

2.4.2.1 Prostate cancer data

For variable selection methods in our study design, we use the 10-fold outer cross validation to get information on variations in the final models of 10 training subsets and test them on each independent test data set. The STMC algorithm uses the 10 replicated 5-fold inner CV for correcting overoptimism. Models are fitted using the Cox proportional hazards regression, and no interaction or nonlinearity effect terms are assumed, and thus, none of them are incorporated in the full model and other models as well. For the comparison of models and selection methods, we achieve the generalization measures of predictive performance based on i) the *C*-index, ii) the integrated AUC (IAUC), iii) the slope of prognostic index, and iv) the integrated Brier score (IBS).

Table 2.5 Comparative analysis for the performance of model selection methods on prostate cancer data and renal transplant data. Model selection methods are Stepwise LRT (likelihood ratio test), Stepwise AIC (Akaike Information Criterion), lasso, and STMC (Stepwise Tuning in Maximum *C*-index).

Dataset	**Prostate Cancer Data**			
Model Selection	LRT	AIC	Lasso	STMC
Train. EPV	21.5	21.5	21.5	21.5
Exp. Model Size	4.8	5.3	6.2	5.1
C-index	0.8043(0.05)	0.8126(0.06)	0.8100(0.04)	**0.8147**(0.06)
IAUC	0.8388(0.05)	0.8532(0.06)	0.8499(0.04)	**0.8552**(0.06)
PI Slope	**0.9606**(0.03)	0.9542(0.05)	0.9598(0.04)	**0.9606**(0.04)
IBS	0.1418(0.02)	0.1390(0.04)	**0.1339**(0.03)	0.1375(0.04)

Dataset	**Renal Transplant Data**			
Model Selection	LRT	AIC	Lasso	STMC
Train. EPV	101	101	101	101
Exp. Model Size	7.8	10.1	12.6	7
C-index	0.6763(0.04)	0.6732(0.04)	0.6768(0.01)	**0.6772**(0.04)
IAUC	0.7312(0.01)	0.7284(0.02)	0.7323(0.02)	**0.7356**(0.03)
PI Slope	0.9589(0.03)	0.9530(0.03)	0.9600(0.03)	**0.9635**(0.04)
IBS	0.0970(0.01)	0.0970(0.01)	**0.0965**(0.01)	0.0968(0.01)

Table 2.5 shows a comparative analysis for the performance of model selection methods on prostate cancer data. The average EPV of training data sets in each method is 21.5, which is greater than 10 and is enough for eluding estimation bias, and the expected model size of each method is 4.8, 5.3, 6.2 and 5.1 for LRT (likelihood ratio test), AIC, lasso, and STMC, respectively. All of the performance measures of STMC are modestly the best on the *C*-index, IAUC, and PI slope. Lasso has the best score only on the IBS and has a largest expected model size. A distribution of predictors, as a result of STMC, yields variable ranking in (a) of Figure. 2.2. In particular, the variables of psa, ece, svi, pgx, and sm have a full frequency of 10, whereas experience is not selected at all.

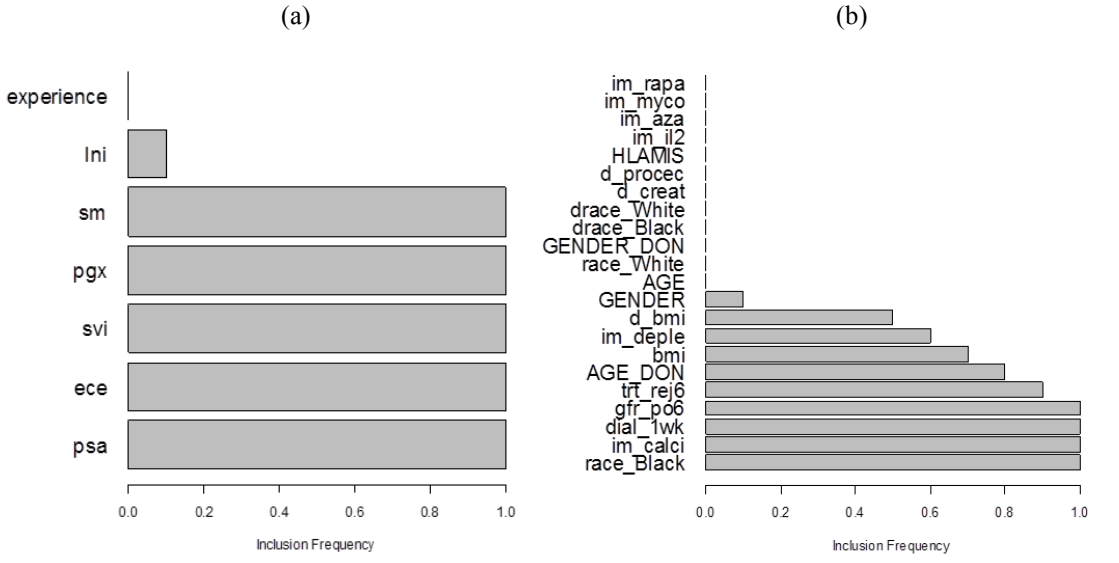


Figure 2.2 Variable ranking of model distribution using STMC (Stepwise Tuning in Maximum *C*-index) in (a) prostate cancer data and (b) renal transplant data.

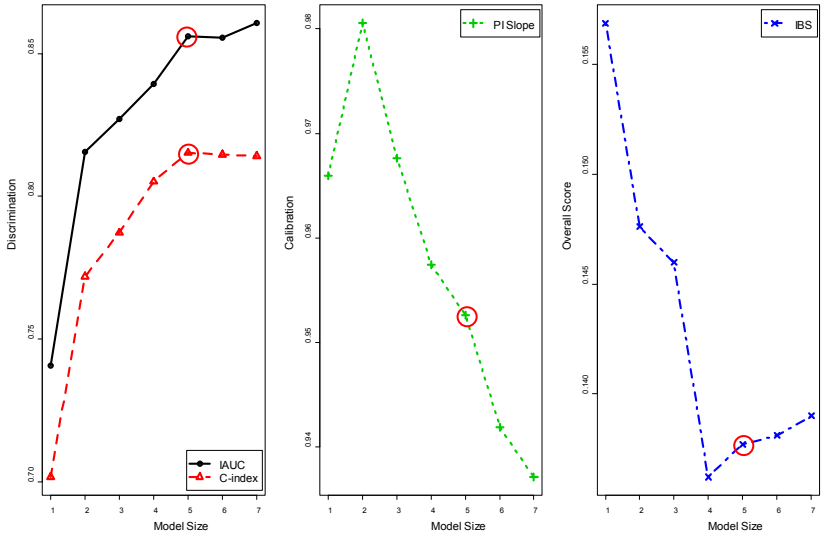


Figure 2.3 Optimization Path of FNSS (Forward Nested Subset Selection) on prostate cancer data.

With FNSS, the optimization path based on *C*-index is displayed in Figure. 2.3 and the results of the other measures are also shown with the corresponding values of discrimination, calibration, and overall score. The open red circle indicates the performance measures of the final model optimized with maximum *C*-index in FNSS. Note that the IAUC has nearly a comparable pattern with the *C*-index, but the IAUC increases and reach to the full model of 7 variables as opposed to the decrease of the *C*-index. We can examine this result numerically in Table 2.6, where the full model and the final model of FNSS are assessed and compared with four primary measures, their development EPV, and the model size computed by 10 CV. The EPV of FNSS grows from 21.5 to 30.04 by the reduction of the model size from 7 to 5. Except for the IAUC,

most measures of the FNSS model are improved as reflected by increases in the *C*-index and the PI slope, and by decreases in the IBS. Table 2.3 also shows the estimated coefficients and statistical significance of predictors in the multivariable Cox proportional hazards model fitted to the entire data on the prostate cancer data for the full model and the final model constructed by FNSS. All of the variables in the FNSS model have significant p-values at the 0.05 level.

Table 2.6 Assessment and comparison of the full model and the final model of FNSS (Forward Nested Subset Selection)

Dataset	Prostate Cancer Data		Renal Transplant Data	
	Full Model	FNSS	Full Model	FNSS
Train. EPV	21.5	30.04	101	329
Expected Model Size	7	5	22	7
C-index	0.8141(0.063)	**0.8153**(0.057)	0.6742(0.040)	**0.6857**(0.039)
IAUC	**0.8607**(0.060)	0.8560(0.057)	0.7387(0.028)	**0.7515**(0.018)
PI Slope	0.9371(0.011)	**0.9567**(0.013)	0.9223(0.059)	**0.9889**(0.039)
IBS	0.1390(0.035)	**0.1377**(0.035)	**0.0965**(0.010)	0.0966(0.107)

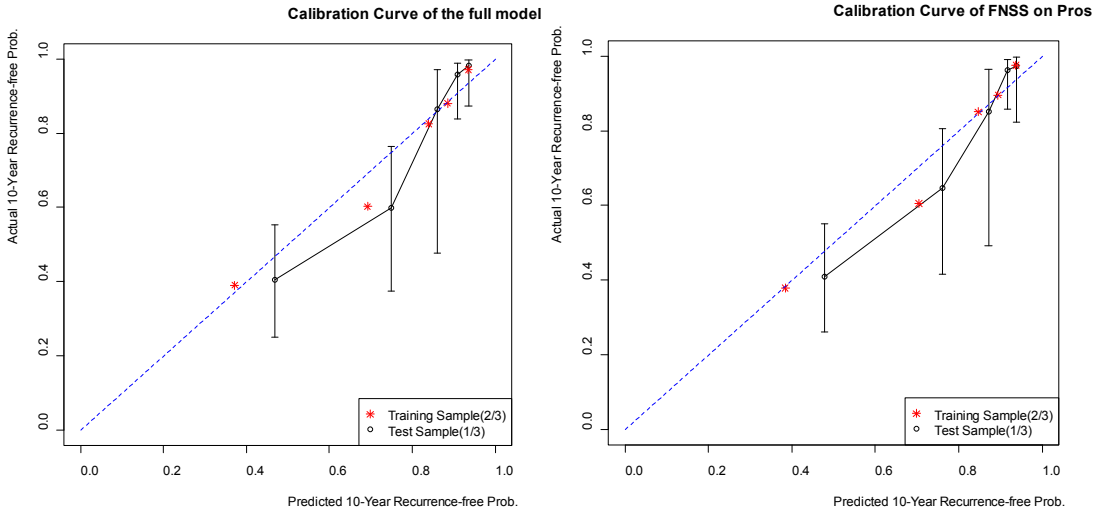


Figure 2.4 Calibration curves of the full model and the final model of FNSS (Forward Nested Subset Selection) on prostate cancer data

The calibration curves of the full model and the FNSS model appear in Figure 2.4. In order to plot the calibration curves of the actual vs. predicted 10-year recurrence-free probability for internal validation, we use a splitting technique using two thirds of the whole data for a training set and one third for a test set, where red asterisks represent the apparent calibration accuracy. In two plots, there seems to be no significant difference in training and test samples, suggesting that it appears to be acceptable in re-substitution results, but is a bit biased in test samples. Note that the results of the plots are based on the specific follow-up-time of 10 years.

2.4.2.2 Renal transplantation data

The same comparative and parameter schemes as used in the prostate cancer data were applied to the renal transplant data. Table 2.5 provides performance measures of the variable selection methods of the LRT, AIC, and STMC on the data set. The development EPV is 101 for all methods alike and the average model size of every method is reduced to approximately half of the full model and the lasso produces a relatively somewhat complex model with 12.6. As with the prostate cancer data, STMC illustrates better performance in all measures except for the IBS, whose value is the largest for the lasso.

The results from the comparative scheme procedure of STMC approximately yield the inclusion frequencies of predictors in the final model in (b) of Figure 2.2. Only 10 variables are included in model selection and the remaining 12 have zero proportions

with no information of ranking.

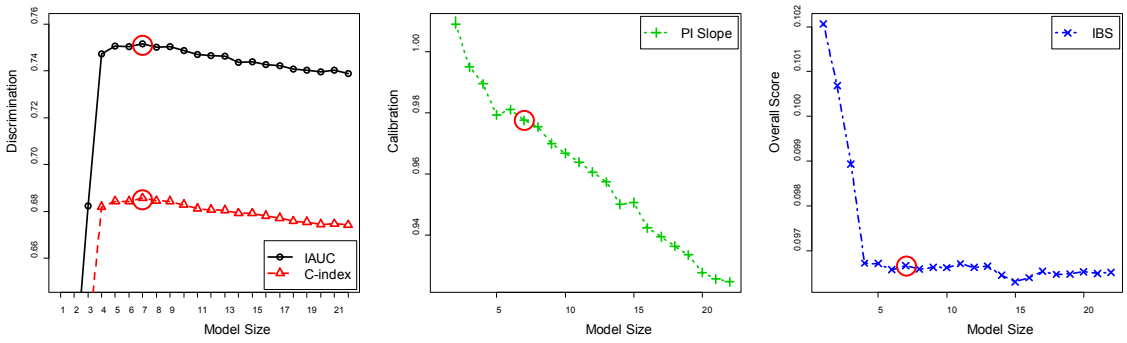


Figure 2.5 Optimization path of FNSS (Forward Nested Subset Selection) on renal transplant data.

The optimization path of Figure 2.5 shows the peaks on the same place of the model size 7 in the IAUC and the *C*-index, whose paths have the similar pattern. The PI slope tends to decrease as the model size increases and the IBS illustrates the sensitivity of the values to the increasing model size.

As shown in Table 2.6, the final model constructed from the FNSS consists of 7 predictors, which are markedly reduced from the 22 variables in the full model and, consequently, the EPV augments from 101 to 329. The FNSS model is improved for every measure, except for the IBS with the difference of 0.0001 from the full model. In the plot of Figure 2.6, the calibration of the prior four groups in the FNSS model is smoother than that in the full model whereas the last group suffers from overestimation bias.

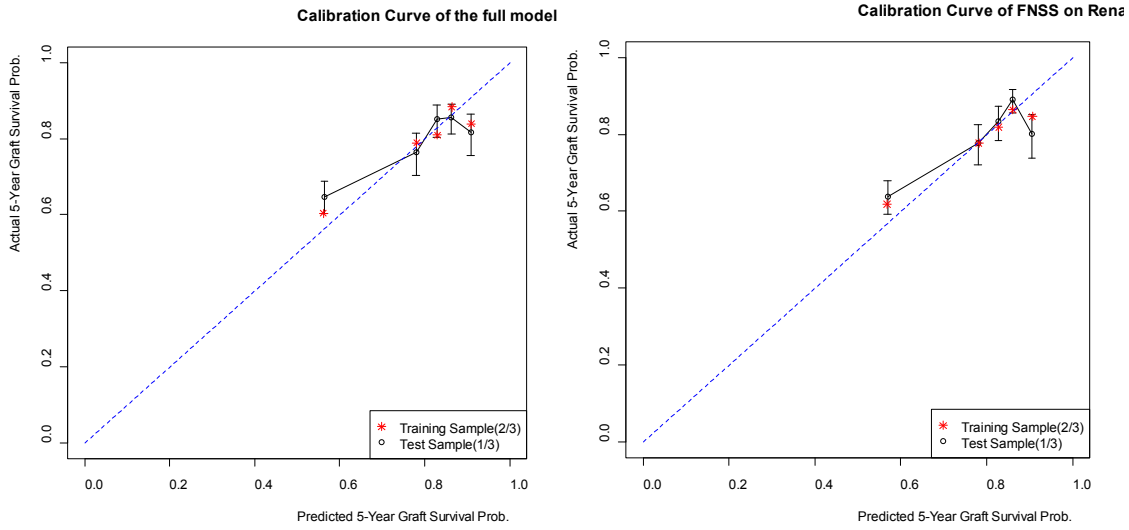


Figure 2.6 Calibration curves of the full model and the FNSS (Forward Nested Subset Selection) model on renal transplant data.

Finally, in Table 2.4 of the multivariable analysis of the full model and the FNSS model, we can see that two p-values (recipient's bmi and donor's age) change significantly and conversely with the decrease of 0.525 to 0.034 and with the increase of 0.02 to 0.978, respectively, and this illustrates the selection bias due to multi-collinearity between the variables that are included and excluded.

2.5. A simulation study

In addition to two clinicopathologic datasets, we present a simulation study to demonstrate the effectiveness of the proposed approach in building a predictive model

with simplified important predictors. We simulate independent clinicopathologic data with p=30 variables and n=2000 individuals. All predictor values are generated from the uniform distribution [0, 1] and the true prognostic index of a linear risk score function $f(\mathbf{x})$ is formed by the coefficients of (10, -10, 7, -7, 3, -3, 1, -1, 0.3, -0.3) for $(x_1, x_2, \dots, x_{10})^{\mathrm{T}}$. The remaining 20 variables of x_{11}-x_{30} are not related to risk time to events. The survival time T is generated from an exponential distribution with parameter $\exp(f(\mathbf{x}))$ when $\mathbf{x}$ is given, and the censoring variable C is generated from an exponential distribution with parameter 0.4. Then we obtain the survival data, $\{(t_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i)) | i = 1, \dots, n\}$ with approximately 50 % of right censoring. We assume that the 30 variables represent the full model. Using multivariate analysis, the variables x_1-x_8, which is related to the time to event, and the randomly generated variables of x_{17}, x_{19}, and x_{30} are significant at the level of 0.05. Our objective is to find a reduced model with variables more relevant to survival information that is at least as accurate as the full model. Table 2.7 shows the results of a simulation study that compares the performance of the different model selection methods and final models. The 10 EPV rule for each model in the development step is held as seen in Table 2.7.

The experimental results demonstrate that the STMC moderately outperforms the LRT and AIC on most measures except the PI slope and is better than the LASSO except for the IBS measure. The LASSO has the poorest performance on the C-index, IAUC, and PI slope. For the expected model size, the LRT is over-simplified with 7.6, and the LASSO built many complex models (of average size 18) that often include many of the

insignificant and non-relevant variables in x_{11}-x_{30} (not shown). AIC models contain variables with the size close to the true model but they are also inconsistently insignificant. STMC models always include the variables x_1-x_8 and many of them have the size of 8. The performance of the final model achieved by FNSS is also moderately better than the full model and true model, and is reduced to a simple model with the 8 variables of x_1-x_8 although it does not include the variables of x_9 and x_{10} that are not significant but related to survival function. Those variables seem to improve each score.

Table 2.7 A simulation study for the performance evaluation of model selection methods: LRT (likelihood ratio test), AIC (Akaike Information Criterion), lasso, and STMC (Stepwise Tuning in Maximum *C*-index), and the final models: the full model, the true model, and FNSS (Forward Nested Subset Selection).

Comparison	Method			
Model Selection	LRT	AIC	LASSO	STMC
Train. EPV	127.24	94.06	31.44	105.43
Exp. Model Size	7.6	10.2	18	9.1
C-index	0.6986(0.01)	0.6957(0.01)	0.6933(0.03)	**0.7016**(0.01)
IAUC	0.7481(0.02)	0.7462(0.02)	0.7460(0.04)	**0.7530**(0.02)
PI Slope	**0.8825**(0.04)	0.8773(0.04)	0.8771(0.05)	0.8790(0.10)
IBS	0.1661(0.04)	0.1663(0.04)	**0.1648**(0.02)	0.1651(0.04)

Comparison	Final model		
Model Selection	Full Model (x_1-x_{30})	True Model (x_1-x_{10})	FNSS (x_1-x_8)
Train. EPV	31.98	95.94	119.93
Exp. Model Size	30	10	8
C-index	0.6898(0.02)	0.7023(0.01)	**0.7025**(0.01)
IAUC	0.7401(0.02)	0.7485(0.02)	**0.7537**(0.02)
PI Slope	0.7943(0.04)	0.8356(0.06)	**0.8620**(0.02)
IBS	0.1674(0.03)	0.1663(0.03)	**0.1659**(0.03)

2.6 Discussion

Our specific goal of this study is to design a model selection method that identifies

a computational model that is optimally reduced based on the *C*-index. To achieve this goal, we have presented the new approach of STMC and FNSS. Using the STMC within 10-fold outer cross validation, we built 10 intermediate optimal final models with the subsets of predictors maximizing the objective criterion, *C*-index and we use the internal validation of 10 replicated 5-fold cross validation for optimism correction. Moreover, instead of using the early stopping strategy that controls the loop number in the STMC, through the optimization path of FNSS, we handle the potential overfitting problem of the STMC stage and the variability of chosen candidate models.

The researchers [56] underline that numerical measures may be difficult to interpret depending on some situations and a model with a good discriminative power will be most relevant for research purposes. Besides, the censoring effect complicates the performance measures of survival models, and we calculate several measures categorized into discrimination (*C*-index, IAUC), calibration (PI Slope and calibration curves), and the overall score (IBS). In particular, the *C*-index is emphasized as a primary accuracy measure of the proposed approach, due to its simplicity and efficiency.

As illustrated in our experiments of two data examples as well as a simulation study using the comparative scheme and the final model of the model selection methods and the final model assessment, the proposed approach demonstrates that the STMC achieves better performance, in the *C*-index, IAUC and PI slope, than other methods. The LASSO method shows good performance on the IBS and yields a relatively complex model. The final model of FNSS, which yields a reduced model, performs better than the

full model in a majority of the measures. On the simulation study, only a random experiment was performed. We need a random experiment design for a comprehensive study.

In the variable selection of survival models, the bias, which may commonly occur in the traditional variable selection methods of a survival data analysis, is i) estimation bias, ii) selection bias, and iii) censoring bias. We discuss, below, what the causes are and how we overcome these kinds of bias in the proposed approach.

When an estimator converges, in probability, to the true parameter as the sample gets larger, it is said to be consistent, which indicates that the estimator is unbiased in large samples. In the Cox regression dealing with the time to event data, the event size is much more essential than the sample size. The partial likelihood of parameters in the Cox regression is maximized, especially over event time, with respect to parameters and they can be estimated by using some version of the Newton-Raphson algorithm. The maximum partial log-likelihood estimation is consistent and unbiased in a sufficiently large event size. Moreover, in connection with the *curse of dimensionality*, a linear increase in the number of variables requires much of the event size geometrically. In general, as the dimension increases, the estimation bias increases when the event size is fixed. This is because many subspaces of features are sparse and empty. Therefore, in survival analysis, the sparse sampling of time to event in high dimensions results in estimation bias in time to event modeling of censored data. Especially, in the literature of survival analysis, a small data is defined by one with less than 10 EPV [42]. Some

references emphasize the ratio of one to ten to prevent the estimation bias. Furthermore, although the EPV is sufficiently large, if there is multicollinearity among independent variables, the parameter estimates remain unbiased but their variances can be large. To somewhat alleviate the problem of the estimation bias, we attempted to correct for the overoptimism using 10 replicated 5-fold cross validation for internal validation [48]. Also, since the estimation under multicollinearity can be unstable in each learning phase of the variable selection method using p-values, this can aggravate the predictive accuracy and may add selection bias during the iteration process. However, in the proposed approach, instead, we employ the strategy to find the final model with a set of predictors optimized for predictive accuracy in the FNSS.

The problem of the censoring bias can be considered via the time dependent measures such as IAUC and IBS, where the information prior to the right censoring of lost to follow-up is used to compute the values over time. Due to the time complexity of the IAUC and the sensitivity to censoring of the IBS, the concordance index is used in our approach for its efficiency, and it tends to have a similar pattern to the IAUC.

Nonlinear or interaction terms can be considered in the multivariable Cox proportional hazards regression to improve model bias but this flexibility may suffer from overfitting. We can handle the problem of the model complexity by the structured model that applies the kernel trick or structured functions giving nonlinear effects [24]. Some authors have proposed multivariable fractional polynomial models using backward elimination with an adaptive algorithm and have compared it with a nonparametric

approach, generalized additive models (GAMs) using cubic smoothing splines [49]. They state that the method can have a risk of overfitting problem and stress that the functional form of the final model should be consistent with medical knowledge. Also, for a more accurate calibration, an appropriate baseline survival function may be required to be specified along with the assumption checking and this may be used in an objective function, which should be less influenced by censoring, for assessing the performance of a model, since discrimination measurement depends on the order of the predicted survival rates.

We have strived to reduce overoptimism using resampling methods in our algorithms and could find a stable set of predictors for the final model. Practically, the prediction of a new patient may still have a variance problem and may be inaccurate and biased for only a single final model. Ensemble methods, which build numerous simpler base models and combine their advantages for a single prediction model, can be designed for the survival model in regression problem.

For the further study, we are investigating the integration of clinicopathologic and genomic data in censored survival analysis. The researchers [5,6] show that, in most of data in the results, the L_2 penalized method (ridge) produced the better performance than the LASSO whose computational cost is very high. However, the ridge method uses the full predictors without parsimony. Furthermore, in the data fusion studies of clinico-genomic data [6], the L_2 penalized method tends to show little improvement in clinical predictors rather than genomic ones. The concern of this research might be to develop a

statistical and computational algorithm that finds the integrative final model with high accuracy and parsimony.

CHAPTER 3 A Hybrid Approach Using Data Integration of Clinicogenomic Information

For the second issue of this dissertation, we contributed to build integrative prognostic models that use clinicopathologic features and predict prognosis after a certain treatment. We introduce a proposed methodology to construct a reduced yet accurate final model with a hybrid signature on high dimensional genomic data with a small sample and investigate competing methods.

3.1 Methods

3.1.1 Cox proportional hazards model for an integrative model

Censored survival data for a combined model is defined by a quartet of variables, $(t_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ for a patient of i, $i \in \{1, .., N\}$. The observed time t_i is given by $t_i = \min(T_i, C_i)$, where T_i is the time of event and C_i is the time of censoring. The event indicator variable, δ_i has a value of 1 if an event occurred at the observed time, or a value of 0 if t_i is censored, and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are the the p_1-vector and p_2-vector of the clinical and molecular variables, respectively.

A proportional hazards integrative model for censored data is defined by

$$h(t|\boldsymbol{x}_i, \boldsymbol{z}_i) = h_0(t) \cdot \exp(CPI_i + GPI_i) \quad (3\text{-}1)$$

where $CPI_i = \boldsymbol{x}_i^T \cdot \boldsymbol{\beta}_C$, and $GPI_i = \boldsymbol{z}_i^T \cdot \boldsymbol{\beta}_G$ are the clinical and genomic prognostic indices, respectively and $h_0(t)$ is a baseline hazard function that is left unspecified with $h_0(t) \geq 0$. To estimate the baseline hazard, we employ the commonly used Breslow estimator written as

$$\hat{h}_0(t_i) = 1/\sum_{t_j \geq t_i} \exp\left(\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_j^{\mathrm{T}} \boldsymbol{\beta}_G\right). \quad (3\text{-}2)$$

The coefficient parameters, $\boldsymbol{\beta}_C$ and $\boldsymbol{\beta}_G$ of the regression are estimated by maximizing the partial log-likelihood (PLL) written as

$$PLL_{full}(\boldsymbol{\beta}_C, \boldsymbol{\beta}_G) = \sum_{i=1}^{N} \delta_i \left[(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\beta}_G) - \ln\left(\sum_{t_j \geq t_i} \exp\left(\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_j^{\mathrm{T}} \boldsymbol{\beta}_G\right)\right)\right]. \quad (3\text{-}3)$$

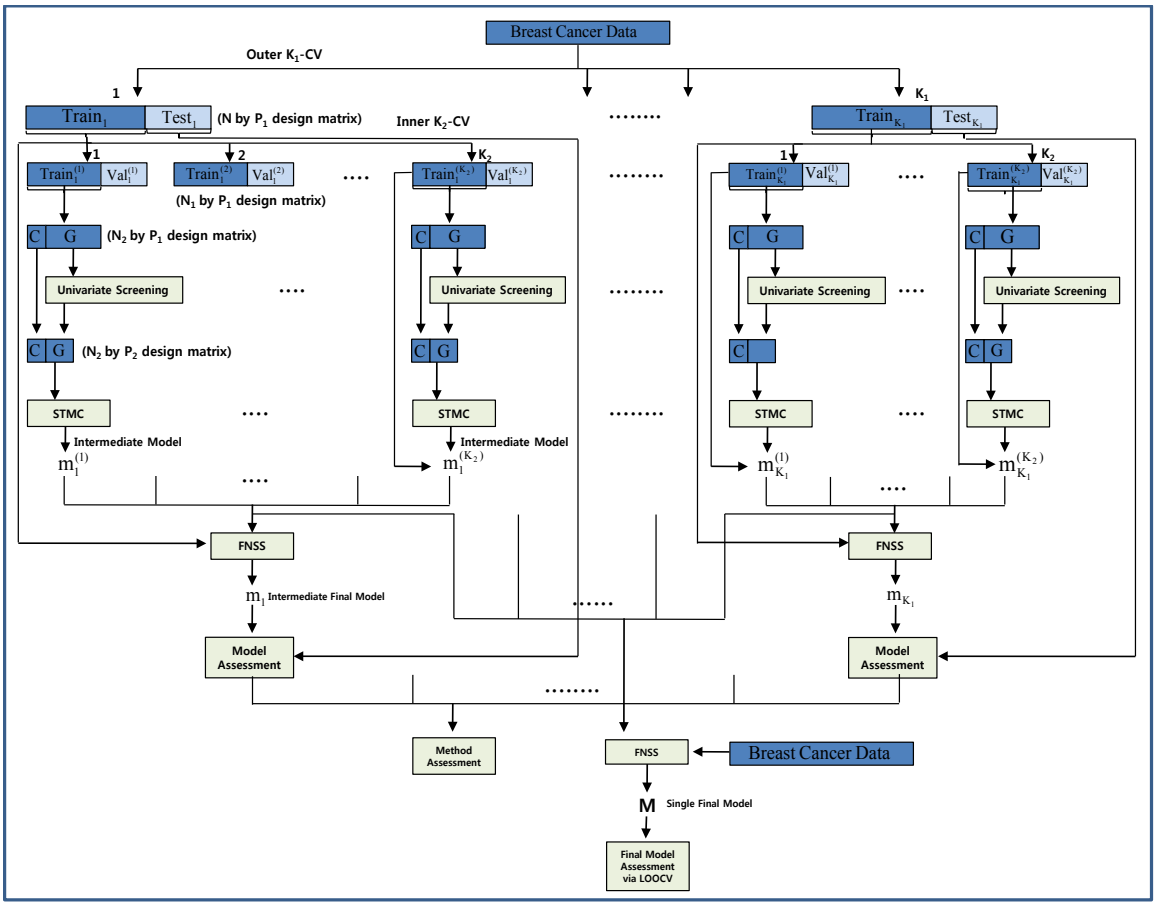


Figure 3.1 A methodology framework to build an integrative model with parsimony.

3.1.2 A methodology framework for integrative model building

Before examining computational algorithms in detail, we briefly present the framework of our proposed integrative modeling methodology. Figure 3.1 displays a methodology framework. All data in outer loop K_1-cross validation (CV) is split into training ($Train_{k_1}$) and test data sets, and the training data is further divided into the

training data ($\text{Train}^{(k_2)}$) and validation data in inner loop K_1-cross validation (CV). $\text{Train}^{(k_2)}$ data are used for developing the model and consist of clinical and genomic variables with an N_2 by P_1 design matrix, in which only genomic variables are presented in the filtering step and a great deal of the dimensionality is reduced in the preliminary univariate screening process using a permutation test based on the concordance index (*C*-index) criterion. The combined and reduced data (N_2 by P_2 design matrix) is fed into the STMC to search an optimized model. However, if $N_2 < P_2$, then the maximum partial log-likelihood estimation of learning using the standard Cox model has infinite solutions leading to the perfect performance. In this case, we first, perform a dimension reduction, using QR factorization prior to the regression. Then maximum partial log-likelihood estimation is applied to fit the data and we obtain an optimized intermediate model. After the results of the inner loop K_2-CV, these K_2 intermediate models are used in a Forward Nested Subset Selection (FNSS) to control overfitting [11]. The difference from the previous version is that the extended version of the FNSS in this study incorporates QR factorization and L_2 penalized Cox regression for the $N << P$ problem but only genomic variables are penalized to adjust for overoptimism among them. The optimal tuning parameter λ^* is selected by using the 10-folds Cross Validated *C*-index (CVCI) for N_λ tuning parameters. Each of the intermediate final models of the outer loop K_1-CV is evaluated using an independent test data set for model assessment. Note that the validation data in the inner loop CV are not used in the STMC but is utilized in the FNSS combined with $\text{Train}^{(k_2)}$ to find the optimized model. Methods are compared based on

several metrics (see Section 2.1 and 3.2.1). The single final model is constructed at the second level of the FNSS using the ($K_1 \times K_2$) intermediate models and is evaluated through LOOCV (see Section 3.2.3).

3.1.3 Dimension reduction

3.1.3.1 Taxonomy of dimension reduction strategies

The advance in genomic studies and the multiple accumulations of microarray data have led to alteration in the dimension reduction techniques for machine learning approaches so that they now play an important role in the success of learning algorithms when a relatively small sample size along with a large number of irrelevant features causes extremely poor performance on independent data. In this study, we use the term "feature" to refer to both clinicopathologic and genomic expression measurements. Dimension reduction techniques can be largely classified as feature extraction or feature selection approaches. Feature extraction methods typically transform the original features of a data set into a reduced number of orthogonal features; examples include Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and QR decomposition, and feature selection algorithms attempt to choose a minimal subset of features while retaining the originals. Furthermore, feature selection techniques can be organized into three categories, depending on how they integrate the feature selection search into learning algorithms for classification and regression: filter methods, wrapper

methods and embedded methods. The description, characteristics, and examples of each approach are summarized well in [20,47]. When used in microarray data analysis, wrapper and embedded approaches generally yield computationally intensive algorithms, although they produce better predictive accurate estimates than the filter approach. Most gene selection algorithms employ filter methods that evaluate genes individually using a t-test, Fisher score, or Wilcoxon rank-sum test [68] for binary outcomes or using Cox proportional hazards modeling for survival outcomes. However, they do not consider the probable correlation information among genes. Hybrid approach to account for the problem was employed in Principal Components Regression (PCR) and Supervised PCR [2] by combining univariate gene selection and feature extraction, and Partial Least Squares (PLS) regression [39] by using components that maximizing the covariance with the outcome in a univariate scheme. Embedded methods, such as the LASSO version of Cox regression [41], are variable selection methods that shrink some of the regression coefficients toward zero by penalizing the coefficient size.

3.1.3.2 Permutation test and preliminary univariate screening

All of the clinical features are included in the initial input, and molecular features are initially filtered using a permutation test described in this subsection, but if the N << P problem still exists, then QR decomposition (Section 3.1.3.3) is further employed to solve the underdetermined system problem.

For preliminary univariate screening, we introduce a model-free permutation test that is more robust than the t-test and more efficient than the Wilcoxon test. Moreover, we can obtain the *P*-value of a metric regarding two quantitative variables or the relationship between each gene and survival outcome information. The preliminary gene selection using the permutation test we proposed is based on the concordance index (*C*-index; Section 2.1) in Cox regression. For each gene, we test the null hypothesis that its gene expression profile is not associated with time to survival or recurrence. Instead of obtaining the null distribution from model assumption, e.g. normal distribution, it is computed by randomly permuting event and nonevent labels that have the same size in developing samples. The *P*-value is the proportion of instances one obtained a *C*-index greater or equal to that of the observed data. After testing genes one by one, we arrange them in order of increasing *P*-value. We then choose the top ranked genes that have a value less than the fixed significance level of 0.05, and these are used as the input in the next step. We compared our proposed preliminary screening using the permutation test with a global test [15] yields a more significant *P*-value, 0.000076 with 459 genes than *P* = 0.98 with 4919 genes in [64] for mortality outcome.

3.1.3.3 Dimensionality reduction using QR decomposition and parameter estimation using space transformation

If the number of combined features is still greater than that of observations (N < P),

then standard multivariate Cox regression yields infinite solutions to estimate parameters, resulting in poor generalization performance, or it simply cannot solve the ill-posed problem due to multicollinearity. To overcome the underdetermined system problem of a design matrix, we take advantage of the geometric intuition that N points in a space of P dimensions lie in an affine subspace of N dimensions and using QR decomposition, project a vector from a higher dimensional space onto some lower dimensional subspace using QR decomposition. The prognostic index N-vector, $\mathrm{PI} = \mathbf{X}\boldsymbol{\beta}$ in Cox regression, where $\mathbf{X}$ is a $\mathrm{N} \times \mathrm{P}$ matrix and $\boldsymbol{\beta}$ is a P-vector, can be factorized into $\mathbf{X} = \mathbf{R}^{\mathrm{T}}\mathbf{Q}_{\mathbf{1}}^{\mathrm{T}}$ using the following QR decomposition.

$$\mathbf{X}^{\mathbf{T}} = \mathbf{Q}\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix} = [\mathbf{Q}_1\mathbf{Q}_2]\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix} = \mathbf{Q}_1\mathbf{R}, \tag{3-4}$$

where $\mathbf{Q}$ is a $\mathrm{P} \times \mathrm{P}$ orthogonal matrix and is partitioned as $\mathbf{Q} = [\mathbf{Q}_1\mathbf{Q}_2]$ with the first N columns and the remaining (P-N) columns of $\mathbf{Q}$, and $\mathbf{R}$ is an $(\mathrm{N} \times \mathrm{N})$ square upper triangular matrix. The columns of $\mathbf{Q}_1$ and $\mathbf{Q}_2$ form an orthonormal basis for span($\mathbf{X}^{\mathbf{T}}$) and for the orthogonal complement of span($\mathbf{X}^{\mathbf{T}}$) respectively, when $\mathbf{X}^{\mathbf{T}}$ has a full column rank. Replacing $\mathbf{X}$ by $\mathbf{R}^{\mathrm{T}}\mathbf{Q}_{\mathbf{1}}^{\mathrm{T}}$ and substituting an N-vector of $\boldsymbol{\theta}$ for $\mathbf{Q}_{\mathbf{1}}^{\mathrm{T}}\boldsymbol{\beta}$, then $\mathrm{PI} = \mathbf{R}^{\mathrm{T}}\mathbf{Q}_{\mathbf{1}}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{R}^{\mathrm{T}}\boldsymbol{\theta}$ and $\boldsymbol{\beta} = \mathbf{Q}_1\boldsymbol{\theta}$. This transformation leads to a data matrix reduction from $\mathbf{X}^{\mathbf{T}}$ to $\mathbf{R}$ and reduces the computational cost from $O(\mathrm{P}^3)$ to $O(\mathrm{PN}^2)$. Figure 3.2 shows how a higher P-dimensional point in columns (observations) of $\mathbf{X}^{\mathrm{T}}$ is projected

to a point with a lower dimensionality of N in columns of **R** by QR decomposition.

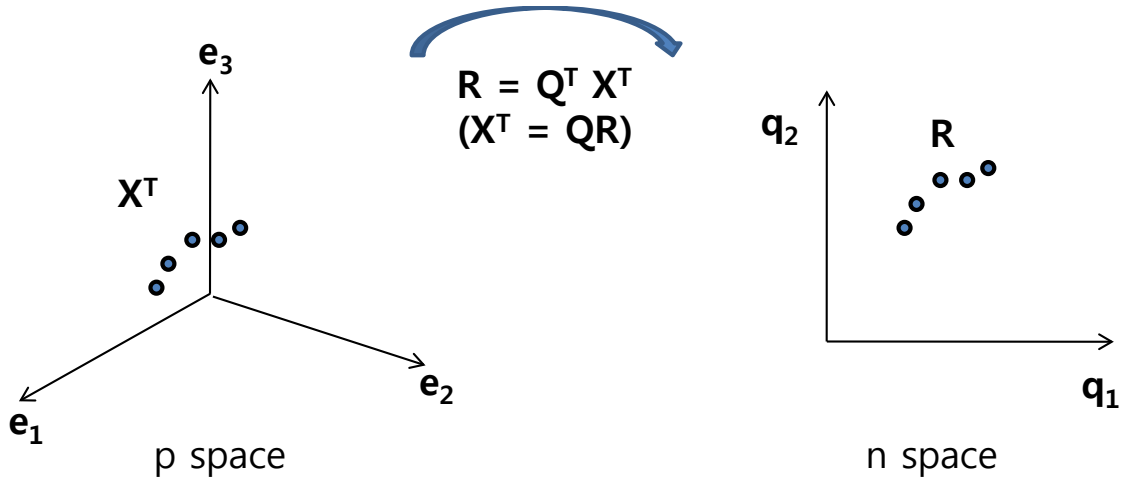


Figure 3.2 Projection from a higher *p*-dimensional point in columns of X^T to a lower *n*-dimensional point in columns

Orthogonal matrices, e.g. $\mathbf{Q}$ ($\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$), have two important properties in geometrical analysis: (1) Euclidean norm preserving property of any vector $\boldsymbol{x}$ between spaces, $\|\mathbf{Q}\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T\mathbf{Q}^T\mathbf{Q}\,\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$, (2) Angle preserving property between any two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$ from P space to N subspace and vice versa; $\langle \mathbf{Q}\boldsymbol{x}, \mathbf{Q}\boldsymbol{y} \rangle = (\mathbf{Q}\boldsymbol{x})^T\mathbf{Q}\boldsymbol{y} = \boldsymbol{x}^T\mathbf{Q}^T\mathbf{Q}\,\boldsymbol{y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$. By these two properties, the Euclidean distance between any two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$ is preserved in two subspaces transformed via an orthogonal matrix and $\mathbf{Q}_1^T$ rotates the columns of $\mathbf{X}^T$ in a coordinate system between subspaces

while reducing the dimensionality from P to N.

3.1.3.4 Extended versions of the STMC and FNSS

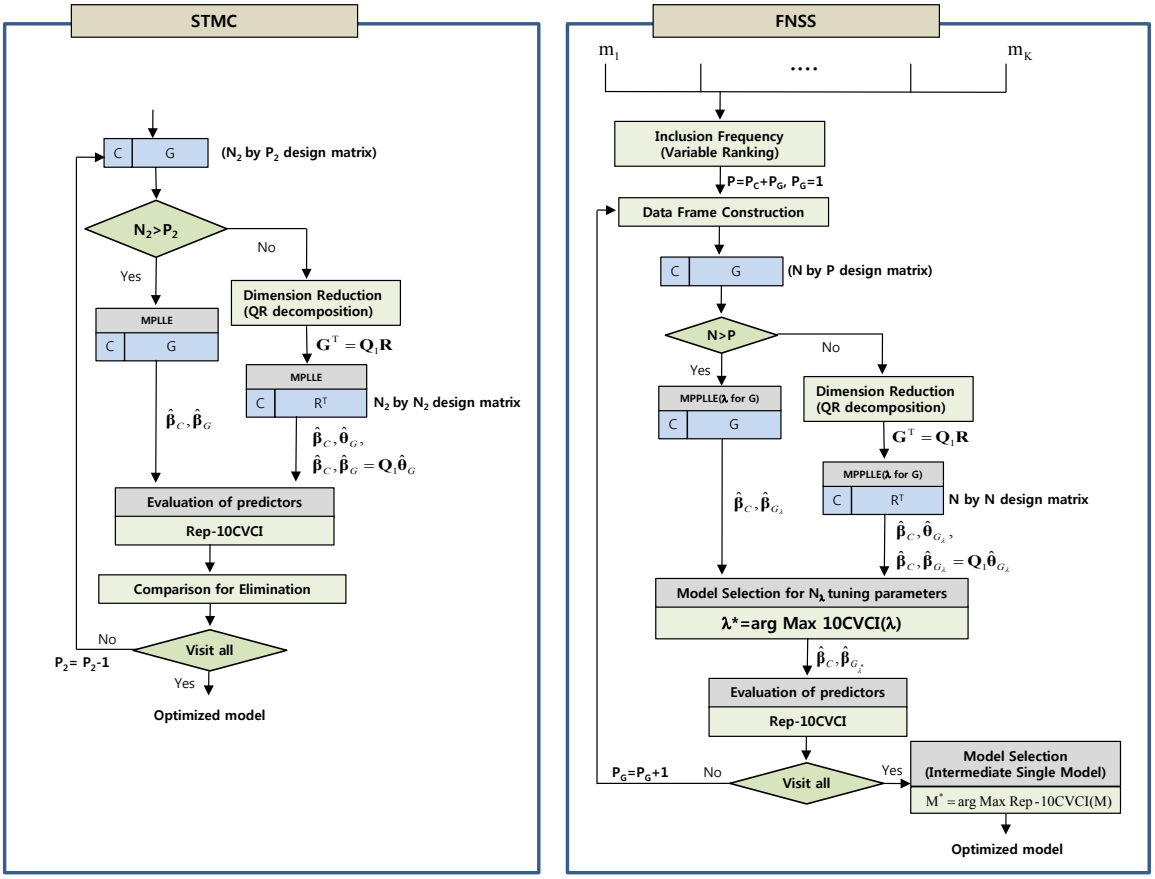


Figure 3.3 Flow diagrams for the extended versions of the STMC and FNSS method.

Flow diagrams of the extended versions of the STMC and FNSS are presented in Figure 3.3. The STMC is a backward stepwise selection method based on the *C*-index in

a Cox model, and thus is a wrapper method [9]. The *C*-index is a metric for evaluating the performance of a group of features in regard to goodness of fit. The iterative process of variable selection performs the optimization search for model complexity in the model space. The algorithm consists of a (1) dropping step, (2) addition step, (3) comparison for choosing an intermediate best model, and 4) stopping rule test. The extended version of the STMC incorporates two techniques such that the current full model is initialized by combined features selected from the preliminary univariate screening step and the dimension reduction using QR decomposition is performed.

The FNSS approach is a filter method [9]. It is designed to control for overfitting resulting from the previous backward stepwise selection and to identify stable single final model. The feature ranking criterion is defined on individual features by the frequency of inclusion in intermediate models produced in inner loop cross validation, which approximately produces the distribution of features of an intermediate final model in an empirical way and is used to interpret their relative relevance. Based on the proportion of inclusion frequency in models, the FNSS builds a model with an increasing set of features while incorporating a single feature one by one from the null model and evaluates the model by the cross-validated *C*-index. The FNSS method selects the subset of features with the maximum cross validated *C*-index as a intermediate final model. The method is also extended to such that L_2 penalization as well as QR decomposition is included in these methods.

Note that only molecular features are applied to the dimension reduction

techniques such as univariate screening and QR decomposition in the STMC and FNSS methods because molecular features suffer from correlation among them and clinical features are stable and have values with small variance.

3.1.4 Competing methods

We compared our proposed method with three others: the forward stepwise selection (FSS), LASSO (L_1 penalization), and ridge regression (L_2 penalization). Currently, ridge Cox regression, which uses L_2 penalized maximum partial log-likelihood estimation, performs better than existing methods [5,6] but it uses too many predictors for analysis and interpretation of survival data. The FSS and LASSO were selected as they are commonly used and well-established. The performance for each method is assessed in the DCV procedure via LOOCV as summarized in Section 3.2.2.

3.1.4.1 Forward stepwise selection (FSS) using the likelihood ratio test (LRT)

Forward stepwise selection starts with the null model and then in our proposed model sequentially includes the feature that has the most significant P-value. The P-value of each feature is computed using the likelihood ratio test (LRT) defined by $2(PLL(\widehat{\boldsymbol{\beta}}) - PLL(\boldsymbol{\beta}^{(0)})$, where $PLL(\widehat{\boldsymbol{\beta}})$ and $PLL(\boldsymbol{\beta}^{(0)})$, which are the log partial likelihood of the

larger model and the nested smaller model, respectively. In order to test the null hypothesis $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$, where $\boldsymbol{\beta}^{(0)}$ are the coefficients of the null model or nested smaller model with p_0 parameters in contrast to the bigger model with the p_1 parameters of $\boldsymbol{\beta}$, the likelihood ratio test is used to derive the significance of a feature for variable selection. The probability distribution of both test statistics is approximated by a chi-square distribution with $(p_1 - p_0)$ degrees of freedom. The strategy used in this method is to sequentially add in the feature yielding the largest value of the likelihood ratio statistic and to stop when no feature yields a statistic greater than a fixed significance level.

3.1.4.2 LASSO (L_1 penalization)

LASSO shrinks the coefficients of a Cox model toward zero by imposing a penalty on their absolute values. It has a built-in feature selection procedure while penalizing the parameters, unlike L_2-penalized Cox regression which uses a quadratic penalty (ridge regression) that allows all coefficients to be non-zero and may yield complex models. The major objective of this shrinkage is to prevent overfitting occurring by collinearity among molecular features. Thus, we fit the parameters $\boldsymbol{\beta}_C$ and $\boldsymbol{\beta}_G$ of clinical features $\boldsymbol{x}_i$ and molecular features $\boldsymbol{z}_i$ respectively for patient i by maximizing L_1 penalized partial log-likelihood (PPLL) defined over the entire data with LASSO penalty λ_G only on $\boldsymbol{\beta}_G$ as follows.

$$PPLL_{\lambda_G}(\boldsymbol{\beta}_C, \boldsymbol{\beta}_G) = PLL_{full}(\boldsymbol{\beta}_C, \boldsymbol{\beta}_G) - \lambda_G \|\boldsymbol{\beta}_G\|_1 \quad (3\text{-}5)$$

$$= \sum_{i=1}^{N} \delta_i \left[(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\beta}_G) - \ln \left(\sum\nolimits_{t_j \geq t_i} \exp(\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_j^{\mathrm{T}} \boldsymbol{\beta}_G) \right) \right] - \lambda_G \|\boldsymbol{\beta}_G\|_1.$$

where $\lambda_G > 0$ and $\|\cdot\|_1$ stands for the L_1 norm. The zero value of λ_G means no shrinkage and the infinity value indicates infinite shrinkage. The cross-validated partial log-likelihood (CVPLL) is used for optimizing tuning parameters.

3.1.4.3 Principal component regression

A principal component regression (PCR) approach in integrative model building is only applied to molecular features $\boldsymbol{Z}_j$, $j = 1,\ldots, P_2$ and produces a small number of linear combinations, principal component ($\mathbf{PC}_m$), $m = 1,\ldots, M$ of $\boldsymbol{Z}_j$, and the $\mathbf{PC}_m$ are then used in place of the $\boldsymbol{Z}_j$ as inputs in integrative Cox regression. We fit the parameters $\boldsymbol{\beta}_C$ and $\boldsymbol{\beta}_{M_G}^{PCR}$ of clinical features $\boldsymbol{x}_i$ and molecular features $\boldsymbol{z}_i$ respectively for patient i by maximizing partial log-likelihood (PLL) for PCR defined with a tuning parameter M_G only for molecular features as follows.

$$PLL_{M_G}^{PCR}\left(\boldsymbol{\beta}_C, \boldsymbol{\beta}_{M_G}^{PCR}\right) \quad (3\text{-}6)$$

$$= \sum_{i=1}^{N} \delta_i \left[(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\beta}_{M_G}^{PCR}) - \ln \left(\sum\nolimits_{t_j \geq t_i} \exp\left(\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{z}_j^{\mathrm{T}} \boldsymbol{\beta}_{M_G}^{PCR} \right) \right) \right],$$

,where $\boldsymbol{\beta}_{M_G}^{PCR} = \sum_{m=1}^{M_G} \theta_m \mathbf{V}_m$, and $\theta_m = \langle PC_m, \boldsymbol{GPI} = \boldsymbol{Z}\boldsymbol{\beta}_{M_G}^{PCR} \rangle / \langle PC_m, PC_m \rangle$ and $\mathbf{V}_m$ is the mth eigenvector of $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$. Actually, $\boldsymbol{GPI} = \sum_{m=1}^{M_G} \theta_m \cdot \boldsymbol{PC}_m$ is used in $PLL_{M_G}^{PCR}$ to estimate the θ_m and the tuning parameter M_G is also determined by using the CVPLL.

3.1.4.4 L_2 penalized maximum partial log-likelihood estimation for ridge Cox regression and the proposed approach

Although the strategy of dimension reduction lessens the computational burden from P to N, highly correlated molecular gene expression measurements still exist. We thus employ the L_2 penalized maximum partial log-likelihood estimation in Cox regression to handle this problem.

We fit the parameters $\boldsymbol{\beta}_C$ and $\boldsymbol{\theta}_G$ of clinical features $\boldsymbol{x}_i$ and genomic features $\boldsymbol{r}_i$ (the columns of **R)** for patient *i* derived from $\mathbf{Z}^{\mathrm{T}} = \mathbf{Q}_1\mathbf{R}$ and $\boldsymbol{\theta}_G = \mathbf{Q}_1^{\mathrm{T}}\boldsymbol{\beta}_G$ using QR decomposition when the number of features is greater than that of observations, by maximizing the L_2 penalized partial log-likelihood (PPLL) defined as below over the data with a quadratic penalty λ_G only on $\boldsymbol{\theta}_G$.

$$PPLL_{\lambda_G}(\boldsymbol{\beta}_C, \boldsymbol{\theta}_G) = PLL_{full}(\boldsymbol{\beta}_C, \boldsymbol{\theta}_G) - \frac{1}{2}\lambda_G \boldsymbol{\theta}_G^{\mathrm{T}} \boldsymbol{\theta}_G \qquad (3\text{-}7)$$

$$= \sum_{i=1}^{N} \delta_i \left[(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{r}_i^{\mathrm{T}} \boldsymbol{\theta}_G) - \ln \left(\sum_{t_j \geq t_i} \exp\left(\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_C + \boldsymbol{r}_j^{\mathrm{T}} \boldsymbol{\theta}_G \right) \right) \right] - \frac{1}{2}\lambda_G \boldsymbol{\theta}_G^{\mathrm{T}} \boldsymbol{\theta}_G$$

The estimate $\widehat{\boldsymbol{\theta}}_G$ is then transformed back to $\widehat{\boldsymbol{\beta}}_G$ through a matrix multiplication as $\widehat{\boldsymbol{\beta}}_G = \mathbf{Q}_1 \widehat{\boldsymbol{\theta}}_G$. Note that the L_2 penalty is invariant under rotations via the orthogonal matrix and two Cox regressions performed in different subspaces are equivalent. The cross-validated partial log-likelihood (CVPLL) and the cross validated *C*-index (CVCI) are used for optimizing tuning parameters for ridge Cox regression and our proposed approach respectively.

3.2 Experimental design

In this section, we describe experimental design on how we will compare the relative performance of the prediction methods independently and how the final single prognostic model is constructed. First, we define the model assessment criteria for the difference in deviance (DD), the *C*-index, the integrated area under the receiver operating curve (IAUC), and the integrated Brier score (IBS) to compare model building methods. Using double cross-validation (DCV), the comparison will be carried out on intermediate final models constructed with each method. The single final model with a hybrid signature will be assessed through leave-one-out cross validation (LOOCV) based on the

log rank test, the *C*-index, and the PI (Prognostic Index) slope and its p-value.

3.2.1 Performance assessment for method comparison

The performance metrics for the predictive accuracy of a prognostic model on censored time to event data are primarily categorized as: (1) *Discrimination*, which measures how well the prediction model can discriminate between cases with events and those without events, and includes the *C*-index and the IAUC, (2) *Calibration*, which quantifies how close a predicted estimate is to the real probability (e.g., the PI slope), (3) *Overall score*, which computes the explained variation in the goodness of the fit, and involves the IBS [56]. In addition to the above general categories, we add the metric, difference in deviance (DD) as the measure of prediction error [5,6]. Here, we define only the DD and the other metrics of the *C*-index, the IAUC, and the IBS are well summalized in Section 2.1.

3.2.1.1 Difference in deviance (DD)

The difference in deviance (DD) between a fitted penalized model and the null model is a measure of a prediction error when evaluating how well a prediction model performs on a test data set. For assessment of our model, we define it on the hybrid features of clinical and genomic features by

$$\mathrm{DD}\left(\widehat{\boldsymbol{\beta}}_C, \widehat{\boldsymbol{\beta}}_{\lambda_G}\right) = -2\left(PPLL^{(\mathrm{test})}\left(\widehat{\boldsymbol{\beta}}_C, \widehat{\boldsymbol{\beta}}_{\lambda_G}\right) - PPLL^{(\mathrm{test})}(\mathbf{0}_C, \mathbf{0}_G)\right), \quad (3\text{-}8)$$

where $PPLL_{full}^{(\mathrm{test})}\left(\widehat{\boldsymbol{\beta}}_C, \widehat{\boldsymbol{\beta}}_{\lambda_G}\right)$ and $PPLL_{full}^{(\mathrm{test})}(\mathbf{0}_C, \mathbf{0}_G)$ are the partial log-likelihood for the test data set evaluated by $\left(\widehat{\boldsymbol{\beta}}_C, \widehat{\boldsymbol{\beta}}_{\lambda_G}\right)$ and $(\mathbf{0}_C, \mathbf{0}_G)$ respectively, where $\mathbf{0}$ indicates a vector of zeros.

3.2.2 Double cross validation (DCV) for comparison of methods

The relative performance measurement of a model for model selection and method comparison are subject to the variability of training samples on account of the EPV, multicollinearity, and right censoring of survival data. Therefore, we need an unbiased estimate of the true performance of the method used to build the integrative model. This can be achieved using K-fold CV because all the aspects of model development, such as model selection and parameter tuning, take place in the training sets within the CV. Consequently, we employ the K_1-fold CV as an outer loop for the assessment of a model building method and a nested K_2-fold CV of training folds within the outer CV to correct for overoptimism. Each of the K_1 intermediate final models during K_1-fold CV is tested on the independent test set left out for evaluation and they are averaged for the assessment of each method.

3.2.2.1 A modified version of the DCV for the proposed approach

In [9], using the DCV, the relative generalization performance for the STMC method was calculated and assessed independently for method comparison. After the FNSS step using the DCV, a single final model was built but the performance test was not independent in order to compare methods. To achieve a method that is independent for comparison, the DCV procedure described above is modified for our proposed approach as in Figure 3.1 and Table 3.1.

Table 3.1 Summary of the procedure for the modified version of DCV.

(1) In outer K_1-fold CV, the entire data is divided into a training (Train_{k_1}) and test data set (Test_{k_1}). $\{k_1 \text{ is an integer } | 1 \le k_1 \le K_1\}$

(2) In inner K_2-fold CV, Train_{k_1} is again split into a training ($\mathrm{Train}_{k_1}^{(k_2)}$) and validation set ($\mathrm{Val}_{k_1}^{(k_2)}$). $\{k_2 \text{ is an integer } | 1 \le k_2 \le K_2\}$

- At each inner fold, the STMC is applied to $\mathrm{Train}_{k_1}^{(k_2)}$ and builds an intermediate model $m_{k_1}^{(k_2)}$.
- The CVCI is used for parameter tuning.
- The purpose of $\mathrm{Val}_{k_1}^{(k_2)}$ is to vary the samples for the generalization performance and the data set is utilized in the first level of FNSS along with $\mathrm{Train}_{k_1}^{(k_2)}$.

(3) After K_2 repetitions, K_2 intermediate models are constructed for each outer fold.

(4) The first level of the FNSS builds an intermediate final model m_{k_1} using K_2 intermediate models.

(5) The final model fitted to Train_{k_1}, m_{k_1} is independently assessed using Test_{k_1}, which is the assessment of the instance for a result of a method.

(6) By averaging the K_1 test performances, the relative generalization performance of the proposed approach is computed for method comparison.

(7) The second level of the FNSS builds a single final model using all of the intermediate models and breast cancer data, and it is evaluated via LOOCV.

In conjunction with a double cross validation strategy, we can define the cross validated penalized partial log-likelihood (CVPPLL) for tuning the model complexity as

$$CVPPLL(\lambda_G) = \sum_{i=1}^{N_g}(PPLL_{full}(\widehat{\boldsymbol{\beta}}_C^{(-g_i)}, \widehat{\boldsymbol{\beta}}_{\lambda_G}^{(-g_i)}) - PPLL_{full}^{(-g_i)}(\widehat{\boldsymbol{\beta}}_C^{(-g_i)}, \widehat{\boldsymbol{\beta}}_{\lambda_G}^{(-g_i)})), \quad (3\text{-}9)$$

where the first term is the *PPLL* evaluated on the entire training data at $(\widehat{\boldsymbol{\beta}}_C^{(-g_i)}, \widehat{\boldsymbol{\beta}}_{\lambda_G}^{(-g_i)})$ and is estimated by the training set except for the held-out set g_i in the inner N_g-fold CV, and the second term is the *PPLL* evaluated on only the training data set at $(\widehat{\boldsymbol{\beta}}_C^{(-g_i)}, \widehat{\boldsymbol{\beta}}_{\lambda_G}^{(-g_i)})$. This strategy is used for parameter tuning in LASSO and ridge Cox regression methods.

3.2.3 LOOCV for a final model assessment

We obtain a final model on the basis of the 10-fold CVCI from the second level of the FNSS using a complete breast cancer data set. The final model is assessed through LOOCV (leave one-out cross validation) on the basis of the following criteria: the log rank test, the *C*-index, and the PI slope and its *p*-value. The LOOCV effectively adjusts the evaluation measures for overly optimistic values when an experimental sample is a relatively small by assigning each individual observation to the test set and selecting the

remaining observations as a training set.

For a final model assessment, we have manipulated the internal validation strategy because adequate publicly available data are lacking. Therefore, independent validation for generalization performance using external validation is required to more accurately assess the performance of the final model.

The LOOCV *C*-index is defined as $C\text{-index} = \sum_{i,j\in\Omega} \mathbf{1}\{PI_i < PI_j\}/|\Omega|$, where the prognostic index, PI_i is a linear combination of the regression coefficients estimated in a training sample and the values of hybrid features in the test data for an individual *i* and Ω is a set of all pairs of patients $\{i, j\}$ that satisfies one of the following conditions: (1) both of the patients *i* and *j* experienced their events and the event risk score PI_i is greater than PI_j $(PI_i > PI_j)$ or, (2) only patient *i* experienced an event and the PI_i is greater than the PI_j with censored time c_j $(PI_i > PI_j)$.

3.2.3.1 PI Slope and its p-value

Calibration can be examined by using the PI slope, α for survival data. It can be computed by performing a Cox regression with the PI_i for the new data set $\boldsymbol{x}_i$ in the LOOCV, as a single covariate in the Cox proportional hazards model as follows.

$$h(t|PI_i) = h_0(t)\exp(\alpha \cdot PI_i), \tag{3-10}$$

where $PI_i = \widehat{\boldsymbol{\beta}}^{(\mathrm{Train.})^{\mathrm{T}}} \cdot \boldsymbol{x}_i^{(\mathrm{Test})}$. If the PI slope is unity, the regression model is perfectly calibrated. Otherwise, the regression coefficients that are estimated in the training sample reflect underestimation or overestimation. Also we perform a hypothesis test of the null hypothesis ($\beta = 0$) vs. the alternative hypothesis ($\beta \neq 0$) using the likelihood ratio test (LRT) and we use the *P*-value as a criterion for model assessment.

For performance measures and model validation, the discrimination and calibration can be combined in a data analysis. These can provide complementary interpretation for comparative analysis, as the overall calibration score is sensitive to censoring mechanisms.

3.2.3.2 Log-rank test

In clinical trials of cancer treatment, the log rank test is a nonparametric hypothesis test that is usually used to compare the survival distributions of two groups. We assign individuals of a test set ($\boldsymbol{x}_i$) to one of the two groups based on their prognostic index PI_i, either good prognosis group or poor prognosis group; the PI median is used as the cut-off. How well the grouping performs is evaluated using the log rank test and its *p*-value.

3.3 APPLICATION TO BREAST CANCER STUDY

3.3.1 Breast cancer dataset

We performed a computational study using the publicly available Netherlands Cancer Institute (NCI) breast cancer data, which contains 24481 gene expression profiles as measured on spot oligonucleotide arrays and 10 clinical covariates from 295 cases of breast carcinoma. This dataset had previously been used to identify and validate a prognostic gene expression profile defined by a set of 70 genes [62,65] and [64] also used it to develop a prognostic signature for survival outcome prediction. The data used in our current study were obtained from [8] and contains 295 patients with 101 metastasis (34%) and 79 death (27%) events. The microarray data was already normalized and background corrected. All patients included in this dataset were younger than 53 years old at diagnosis and had stage I (tumors $\leq$ 2.0cm) or II disease (tumors $\geq$ 2.0cm). The median follow-up among all 295 patients was 6.8 years (range, 0.05-18.3) for metastasis outcome and 7.2 years (range, 0.05-18.3) for mortality outcome. Multiple missing expression values were imputed by using the K-nearest neighbor estimation method [60] to minimize the effect of incomplete data on analyses.

The clinical data consisted of the following 10 variables, which are of 3 types: (1) the continuous variables are Age (per year), Lymph Node Status (number of positive nodes), and Diameter (per cm); (2) the categorical variables are T1_T2 (tumors $\leq$ 2.0cm vs. tumors $\geq$ 2.0cm), Estrogen Receptor Status (positive vs. negative), Mastectomy (yes vs. no), Chemotherapy (yes vs. no), and Hormonal Therapy (yes vs. no); (3) the ordinal

variables are Vascular Invasion (0 vessels, 1-3 vessels, and >3 vessels), and Tumor Grade (well differentiated, intermediate differentiated, and poorly differentiated). Note that 74 patients of 295 died of metastatic breast cancer and it is 94 percent of patients who died.

3.3.2 Experimental results

3.3.2.1 Metastasis outcome

As a preliminary gene-filtering step, we applied the univariate screening using a permutation test in a modified version of the DCV to each of the training sets for metastasis outcome, in which the 10-fold CV Concordance index was used as a reliable performance result, yielding on average 101 genes with *P*-values < 0.05. Then, this genomic information was used for building the molecular and integrative models. The initial predictor size as an input for each approach was 10, ~102, and ~112 in the clinical, molecular, and integrative models, respectively. Because the molecular and integrative model size is on average smaller than that of the subsamples, model inference was performed using maximum partial log-likelihood estimation (MPLLE) and maximum penalized partial log-likelihood estimation (MPPLLE) without QR decomposition in the STMC and FNSS method, respectively. For feature selection in the STMC and an optimized model in the FNSS, the set of features is evaluated using the twice-replicated 10-fold CV *C*- index for unbiased estimation. In order to tune the MPPLLE parameters in the FNSS, the 10-fold CVPPLL is used for the LASSO and ridge Cox regression

approaches, and the 10-fold CV *C*-index is used for our approach.

In the modified version of the DCV, we assigned (K_1-1) to K_2 so that the same fold positions could be used in both the outer and inner CV, and one fold was reserved for the independent test in the outer CV. Thus, the maximum inclusion frequency in the first level of the FNSS was 9 and $K_1 \cdot K_2/2 = 45$ was used in the second level of the FNSS. The STMC was carried out only for $k_2 > k_1$ and the previously generated model was used to test for $k_2 \leq k_1$. The single parsimonious final model was constructed on the basis of the 45 intermediate models. Note that if there exist tied proportions in the FNSS, the univariate ranking *C*-index score was used to place in order.

Table 3.2 shows the results of the comparative analysis of methods (FSS, LASSO, Ridge, PCR and the proposed) for the clinical, molecular, and integrative models for predicting metastasis in the breast cancer. All methods used univariate screening except the clinical model. Method performances which were the best or on which methods tie are indicated in bold and the smallest expected model size is indicated in italics. In the clinical model, the proposed method performed better on the majority of performance measures as also reported in [9]. For the development of the molecular and integrative models, the proposed methodology outperformed others in all performance measures. Note that in the clinical model, the FSS was the best on the IBS, and in the integrative model, the ridge and proposed methodology tied on the IBS. In regard to model size, the FSS yielded the smallest model except the PCR that uses supergenes, but it performed the worst in all measures in the molecular and integrative model building and in particular on

the DD, revealing that it is worse than the null model (DD=0); however, the proposed method performed better with a few more features. The results for our methodology show that although the clinical model was better than the molecular model alone, the integrative model enhanced the performance of the clinical model.

Table 3.2 Comparative performance analysis of methods for the clinical (10 variables), molecular (~102 variables; *P*-value < 0.05), and integrative model (~112 variables) for breast cancer data (BCD) on the metastasis event. All methods use univariate screening, except the clinical model.

	Dataset	BCD on Metastasis				
	Method	FSS	LASSO	Ridge	PCR	Proposed
Clinical Model	Exp. Model Size	*3.1*(0.32)	7.2(1.55)	10	-	4.9(0.99)
	C-index	0.708(0.10)	0.700(0.04)	0.703(0.05)	-	**0.716**(0.05)
	IAUC	0.760(0.11)	0.757(0.05)	0.762(0.06)	-	**0.763**(0.06)
	IBS	**0.189**(0.03)	0.193(0.02)	0.215(0.06)	-	0.206(0.06)
	DD	-3.440(3.96)	-3.010(3.16)	-2.927(1.20)	-	**-4.277**(2.73)
Molecular Model	Exp. Model Size	12.7(2.54)	33.8(4.92)	102	*2.9*(1.449)	16.1(2.69)
	C-index	0.611(0.05)	0.630(0.07)	0.687(0.07)	0.665(0.097)	**0.697**(0.08)
	IAUC	0.646(0.07)	0.665(0.07)	0.736(0.08)	0.702(0.118)	**0.739**(0.07)
	IBS	0.253(0.05)	0.218(0.03)	0.221(0.06)	0.227(0.022)	**0.214**(0.06)
	DD	22.23(34.76)	2.476(8.70)	-1.227(3.98)	1.535(2.27)	**-3.019**(6.38)
Integrative Model	Exp. Model Size	12.4(1.06)	15(7.44)	112	*12.3*(0.07)	19.8(4.05)
	C-index	0.660(0.07)	0.681(0.08)	0.751(0.06)	0.744(0.04)	**0.759**(0.03)
	IAUC	0.703(0.07)	0.735(0.08)	0.811(0.07)	0.807(0.03)	**0.823**(0.03)
	IBS	0.225(0.06)	0.198(0.03)	**0.188**(0.07)	0.192(0.02)	**0.188**(0.06)
	DD	24.86(64.6)	-1.995(5.09)	-5.665(4.99)	2.739(3.16)	**-7.062(3.05)**

	Dataset	BCD on Metastasis	
	Method	BestRand(1000)	Final Model
Molecular Model	Exp. Model Size	15	15
	C-index	0.593	**0.735**(0.079)
	IAUC	0.596	**0.778**(0.084)
	IBS	0.214	**0.199**(0.074)
	DD	-0.663	**-5.057**(5.442)
Integrative Model	Exp. Model Size	23	23
	C-index	0.631(0.07)	**0.767**(0.07)
	IAUC	0.670(0.08)	**0.827**(0.08)
	IBS	0.217(0.05)	**0.188**(0.07)
	DD	-0.878(1.93)	**-7.48**(5.11)

- The smallest expected model size is indicated in *italics*.
- Method performances which are the best or on which methods tie are indicated in bold.

To further verify the effectiveness of our approach, we define the best random method (BestRand) such that if our algorithm returns a model of size n, we select n features randomly from the original hybrid set of the clinical feature (10) and molecular feature (24481), repeat 1000 times using boostrapping resampling, and select the best set of n random features. Table 3.2 shows that the final model performance of our proposed method in the molecular and integrative model outperform the BestRand method on all measures.

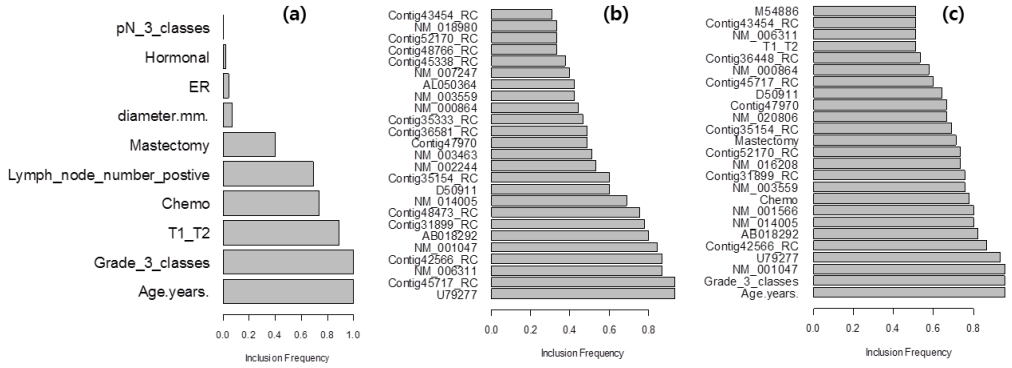


Figure 3.4 Feature Relevance Ranking (FRR) of the model distribution for metastasis obtained from intermediate models of an extended version of STMC. (a) Clinical model, (b) Molecular model, (c) Integrative model.

Figure 3.4 displays the feature relevance ranking (FRR) of the model distribution for metastasis outcome obtained from the 45 intermediate models using an extended version of the STMC in (a) the clinical, (b) molecular, and (c) integrative model. After applying our methodology to build a single final model, we selected a hybrid signature,

as described in Table 3.3, for (a) the clinical model with 5 clinical factors, (b) the molecular model with 15 genomic biomarkers, and (c) the integrative model with 5 clinical factors and 18 genomic biomarkers. Many clincial and molecular features in each model overlapped with those in the integrative model. The clinical risk factors were more dominant than the molecular biomarkers in the FRR for the integrative model, and 7 unidentified genes in the model were included in the molecular features (18). The top set of panels in Figure 3.5 shows the Kaplan-Meier curves, in which patients are devided into the high and low risk groups by using the median PI, and assessment of the single final model via LOOCV using the log rank test, the *C*-index, the PI slope and its P-value in the clinical, molecular, and integrative model. It is not readily apparent whether the clinical or molecular model performs better. However, the integrative model obviously outperformed both of them and it is clear that the metastasis free probailities between the two groups show the steep difference with the small censoring effect from 0 to 5 years.

Table 3.3 The description of features in the single final model built by the proposed approach for metastasis. (a) Clinical model (5 clinical factors), (b) Molecular model (15 genes), (d) Integrative model (5 clinical factors and 18 genes).

(a)

Clinical Var.	FRR	Description	Type
Age.years	**1**	Age (Year)	Cont.
Grade_3_classes	**2**	(Well, intermediately, poorly) differentiated	Ord.
T1_T2	**3**	tumors ≤ 2.0cm vs. tumors ≥ 2.0cm	Ord.
Chemo	**4**	Chemotherapy (Yes vs. No)	Cat.
LNP	**5**	Lymph node number positive (Yes vs. No)	Cat.

(b)

Molecular Var. (Accession #)	FRR	Gene Description	Symbol
U79277	**1**	-	-
Contig45717_RC	**2**	-	-
NM_006311	**3**	nuclear receptor corepressor 1	NCOR1
Contig42566_RC	**4**	-	-
NM_001047	**5**	steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1)	SRD5A1
AB018292	**6**	Homo sapiens mRNA for KIAA0749 protein, partial cds.	KIAA0749
Contig31899_RC	**7**	-	-
Contig48473_RC	**8**	-	-
NM_014005	**9**	protocadherin alpha 9	PCDHA9
D50911	**10**	vestigial like 4 (Drosophila)	KIAA0121
Contig35154_RC	**11**	-	-
NM_002244	**12**	potassium inwardly-rectifying channel, subfamily J, member 12	KCNJN1
NM_003463	13	protein tyrosine phosphatase type IVA, member 1	PTP4A1
Contig47970	**14**	-	-
Contig36581_RC	**15**	-	-

FRR : Feature Relevance Rank

(c)

Clinical Var.	FRR	Description	Type
Age.years	**1**	Age (Year)	Cont.
Grade_3_classes	**2**	(Well, intermediately, poorly) differentiated	Ord.
Chemo	**9**	Chemotherapy (Yes vs. No)	Cat.
Mastectomy	**14**	Mastectomy (Yes vs. No)	Cat.
T1_T2	**22**	tumors ≤ 2.0cm vs. tumors ≥ 2.0cm	Ord.

Molecular Var. (Accession #)	FRR	Gene Description	Symbol
NM_001047	**3**	steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1)	SRD5A1
U79277	**4**	Human clone 23548 mRNA sequence	-
Contig42566_RC	**5**	-	-
AB018292	**6**	Homo sapiens mRNA for KIAA0749 protein, partial cds.	KIAA0749
NM_014005	**7**	protocadherin alpha 9	PCDHA9
NM_001566	**8**	inositol polyphosphate-4-phosphatase, type I, 107kDa	INPP4A
NM_003559	**10**	phosphatidylinositol-5-phosphate 4-kinase, type II, beta	PIP5K2B
Contig31899_RC	**11**	-	-
NM_016208	**12**	vacuolar protein sorting 28 homolog (S. cerevisiae)	LOC51160
Contig52170_RC	**13**	-	-
Contig35154_RC	**15**	-	-
NM_020806	**16**	gephyrin	GPH
Contig47970	**17**	-	-
D50911	**18**	vestigial like 4 (Drosophila)	KIAA0121
Contig45717_RC	**19**	-	-
NM_000864	**20**	5-hydroxytryptamine (serotonin) receptor 1D	HTR1D
Contig36448_RC	**21**	-	-
NM_006311	**23**	nuclear receptor corepressor 1	NCOR1

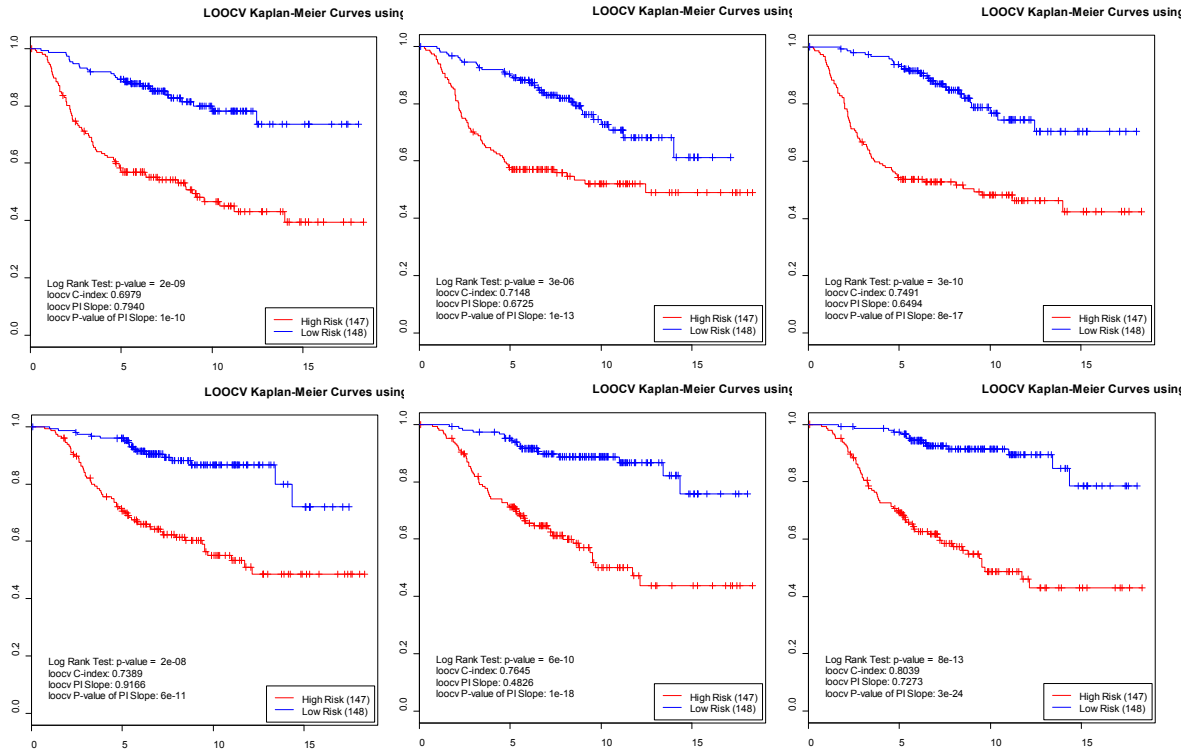


Figure 3.5 The KM curves and the single final model assessment via LOOCV using the log rank test, *C*-index, PI slope and its *P*-value. The three panels of the first row are the results for metastasis outcome and the three panels of the second row are the results for mortality outcome (left column: Clinical model, middle column: Molecular model, right column: Integrative model). The horizontal coordinate of KM curves is the time of year, and the vertical coordinate of the first row is the metastasis free prob. and that of the second row is the survival prob.

3.3.2.2 Mortality outcome

The same model building scheme was applied to mortality outcome. Using the permutation test, the preliminary univariate screening produced on average 459 genes with *P*-values < 0.05. The initial feature size for each method was 10, ~459, and ~469 in the clinical, molecular, integrative model, respectively. In this instance, due to the P >> N

problem, the intermediate models were estimated using QR decomposition in the STMC and FNSS.

Table 3.4 Comparative performance analysis of methods for the clinical (10 variables), molecular (~459 variables; *P*-value < 0.05), and integrative model (~469 variables) for breast cancer data (BCD) on the death event. FSS and lasso use univariate screening, and ridge and the proposed method use univariate screening and QR decomposition, except for the clinical model.

	Dataset	BCD on Mortality				
	Method	FSS	LASSO	Ridge	PCR	Proposed
Clinical Model	Exp. Model Size	*2.6*(0.70)	5.7(0.68)	10	-	4.8(1.05)
	C-index	0.710(0.09)	0.722(0.09)	0.733(0.10)	-	**0.739**(0.10)
	IAUC	0.732(0.13)	0.750(0.13)	0.740(0.13)	-	**0.757**(0.15)
	IBS	0.175(0.04)	**0.173**(0.04)	0.188(0.07)	-	0.187(0.07)
	DD	-3.755(4.06)	-3.624(4.00)	-4.166(4.09)	-	**-4.301**(2.77)
Molecular Model	Exp. Model Size	20(4.38)	29.9(26.08)	459	*2.6*(0.84)	15.9(3.28)
	C-index	0.585(0.07)	0.556(0.11)	0.738(0.12)	0.699(0.107)	**0.769**(0.11)
	IAUC	0.646(0.06)	0.578(0.14)	0.806(0.09)	0.733(0.109)	**0.809**(0.09)
	IBS	0.243(0.08)	0.188(0.05)	0.173(0.08)	0.171(0.043)	**0.165**(0.07)
	DD	32.892(25.90)	1.261(3.33)	-2.729(5.53)	-0.97(2.68)	**-5.496**(5.34)
Integrative Model	Exp. Model Size	15.4(3.27)	26.7(14.24)	469	*12.8*(0.04)	19(1.94)
	C-index	0.692(0.11)	0.741(0.11)	0.769(0.91)	0.775(0.07)	**0.802**(0.10)
	IAUC	0.713(0.13)	0.759(0.15)	0.802(0.12)	0.814(0.08)	**0.826**(0.09)
	IBS	0.197(0.05)	0.191(0.06)	0.181(0.12)	0.163(0.05)	**0.162**(0.07)
	DD	9.441(14.18)	-3.461(4.02)	-5.420(4.11)	-0.814(6.14)	**-7.628**(5.80)

	Dataset	BCD on Mortality	
	Method	BestRand(1000)	Final Model
Molecular Model	Exp. Model Size	16	16
	C-index	0.607	**0.777**(0.075)
	IAUC	0.613	**0.804**(0.077)
	IBS	0.212	**0.176**(0.074)
	DD	-0.925	**-5.64**(1.65)
Integrative Model	Exp. Model Size	17	17
	C-index	0.649(0.09)	**0.821**(0.084)
	IAUC	0.662(0.13)	**0.853**(0.073)
	IBS	0.200(0.07)	**0.153**(0.070)
	DD	-1.261(3.06)	**-9.137**(4.950)

- The smallest expected model size is indicated in *italics*.
- Method performances which are the best or on which methods tie are indicated in bold.

Table 3.4 shows the results of the comparative analysis of methods (FSS, LASSO, ridge , PCR, and the proposed methods) for the clinical, molecular, and integrative model. All methods used univariate screening, and the ridge and proposed method used QR decomposition, which was not used in the clinical model. In the clinical model, the proposed methodology performed better on most measures than the other methods. For the development of the molecular and integrative model, the proposed methodology surpassed others on all performance measures. Note that the LASSO method had the best IBS in the clinical model and that in the molecular model, the FSS and LASSO were overfitted to the training set and performed even worse than in the null model in the test set. Table 3.4 also shows that the final model performance of our proposed method in the molecular and integrative model outperform the BestRand method on all measures.

Although the size of the FSS or PCR was the smallest in all model building, the proposed methodology also had better performance with a few more features. In contrast to our results for the outcome of metastasis, when we used our methodology to build a model for predicting mortality, the molecular model outperformed the clinical model, and the integrative model also enhanced the performance of the molecular model.

Table 3.5 The description of features in the single final model built by the proposed approach for mortality outcome. (a) Clinical model (4 clinical factors), (b) Molecular model (16 genes), (d) Integrative model (4 clinical factors and 13 genes).

(a)

Clinical Var.	FRR	Description	Type
Grade_3_classes	**1**	(Well, intermediately, poorly) differentiated	Cont.
ER	**2**	Estrogen Receptor Status (Pos. vs. Neg.)	Ord.
Age.years	**3**	Age(Year)	Ord.
T1_T2	**4**	tumors ≤ 2.0cm vs. tumors ≥ 2.0cm	Cat.

(b)

Molecular Var. (Accession #)	FRR	Gene Description	Symbol
NM_014918	**1**	chondroitin sulfate synthase 1	KIAA0990
Contig34060_RC	**2**	-	-
NM_016208	**3**	vacuolar protein sorting 28 homolog (S. cerevisiae)	LOC51160
Contig12842_RC	**4**	-	-
NM_003860	**5**	barrier to autointegration factor 1	BCRP1
Contig45717_RC	**6**	-	-
Contig27761_RC	**7**	-	-
NM_000479	**8**	anti-Mullerian hormone	AMH
NM_005267	**9**	gap junction protein, alpha 8, 50kDa	GJA8
D50911	**10**	vestigial like 4 (Drosophila)	KIAA0121
NM_007218	**11**	ring finger protein 139	TRC8
Contig47888_RC	**12**	-	-
NM_001427	13	engrailed homeobox 2	EN2
AB018304	**14**	chloride channel CLIC-like 1	KIAA0761
NM_001232	**15**	calsequestrin 2 (cardiac muscle)	CASQ2
Contig9553_RC	**16**	-	-

FRR : Feature Relevance Rank

(c)

Clinical Var.	FRR	Description	Type
Grade_3_classes	**4**	(Well, intermediately, poorly) differentiated	Ord.
ER	**5**	Estrogen Receptor Status (Pos. vs. Neg.)	Cat.
Age.years	**9**	Age(Year)	Cont.
Chemo	**11**	Chemotherapy (Yes vs. No)	Cat.
Molecular Var. (Accession #)	**FRR**	**Gene Description**	**Symbol**
NM_003860	**1**	barrier to autointegration factor 1	BCRP1
Contig12842_RC	**2**	-	-
Contig34060_RC	**3**	-	-
Contig45717_RC	**6**	-	-
NM_005267	**7**	gap junction protein, alpha 8, 50kDa	GJA8
Contig27761_RC	**8**	-	-
NM_016208	**10**	vacuolar protein sorting 28 homolog (S. cerevisiae)	LOC51160
NM_000479	**12**	anti-Mullerian hormone	AMH
Contig9553_RC	**13**	-	-
D50911	**14**	vestigial like 4 (Drosophila)	KIAA0121
AB018304	**15**	Mid-1-related chloride channel 1 isoform 1	KIAA0761
NM_001427	**16**	engrailed homeobox 2	EN2
Contig47888_RC	**17**	-	-

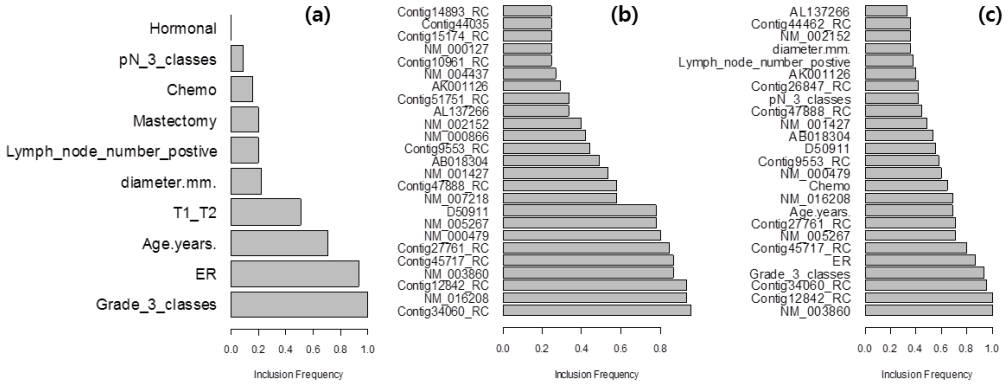


Figure 3.6 Feature Relevance Ranking (FRR) of the model distribution for mortality obtained from intermediate models of an extended version of STMC. (a) Clinical model, (b) Molecular model, (c) Integrative model.

Figure 3.6 displays the feature relevance ranking (FRR) of the model distribution for mortality in (a) the clinical, (b) molecular, and (c) integrative model. After applying our methodology to build a single final model, we selected a subset of features, as described in Table 3.5, for (a) the clinical model with 4 clinical factors, (b) the molecular model with 16 genomic biomarkers, and (c) the integrative model with 5 clinical factors and 18 genomic biomarkers. The estrogen receptor status, ER that was not contained in the metastasis event models was selected in the clinical and integrative model, and it appears that the ER status is the important clinical risk factor in survival anaysis of the breast cancer data set. All of the genomic features in the integrative model were included in the molecular model. The genomic biomarkers were more dominant than the clinical risk factors in the integrative model FRR, and 6 unidentified genes in the model were

included in the genomic features (13). The bottom set of panels in Figure 3.5 shows the Kaplan-Meier curves using the median PI and assessment of the single final model via LOOCV in the clinical, molecular, and integrative model. In analogy to the metastasis outcome, it is not readily discernible as to whether the clinical or molecular model performs better, but the integrative model clearly outperformed both, and it is also apparent that the Kaplan Meier survival probailities differ steeply between the two groups with the small censoring effect from 0 to 5 years.

3.4 Molecular model for DLBCL data

3.4.1 DLBCL data

The survival of patients with diffuse large-B-cell lymphoma (DLBCL) after chemotherapy is influenced by molecular features of the cancers. We used the gene-expression profiles of these lymphomas to develop molecular models of survival. We examined the DLBCL dataset [46] and there were 7399 genes from 240 patients with the use of DNA microarrays. All patents had received anthracycline-based chemotherapy. Median follow-up was 2.8 years overall (7.3 years for survivors), and 57 percent of patients died during this period. The median age of the patients was 63 years, and 56 percent were men.

3.4.2 Experimental results

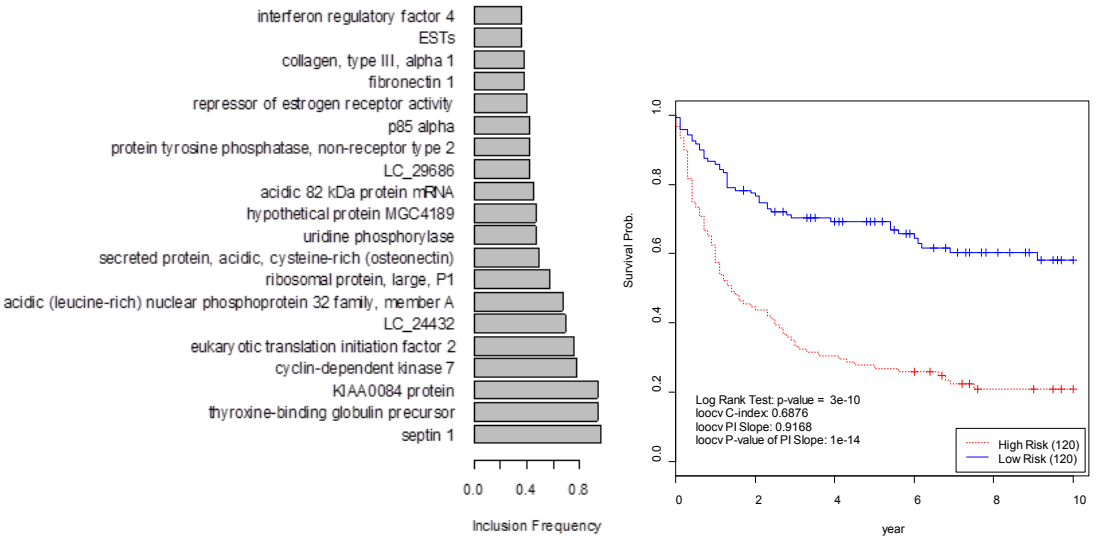
Table 3.6 Comparative performance analysis of methods for the molecular model (~39 variables; *P*-value < 0.05) for DLBCL data on the death event.

	Dataset	**DLBCL Data**				
	Method	FSS	LASSO	Ridge	PCR	Proposed
Molecular Model	Exp. Model Size	3.3(0.675)	7.9(1.66)	39(3.45)	*1.3*(0.675)	8.5(3.54)
	C-index	0.606(0.08)	0.628(0.08)	0.633(0.08)	0.632(0.08)	**0.676**(0.05)
	IAUC	0.647(0.10)	0.665(0.11)	0.668(0.103)	0.674(0.10)	**0.705**(0.08)
	IBS	0.247(0.04)	0.232(0.03)	0.261(0.09)	0.230(0.02)	**0.218**(0.06)
	DD	0.627(5.86)	2.52(02.37)	-1.42(3.11)	0.888(2.47)	**-4.53**(1.82)

- The smallest expected model size is indicated in *italics*.
- Method performances which are the best are indicated in bold.

Molecular prognostic models were constructed with expression patterns that were associated with survival. Several model building methods were applied to the DLBCL dataset using the DCV procedure in the same manner as before. Table 3.6 shows the comparative performance analysis of methods for the molecular model fitted to the DLBCL data for mortality outcome. The preliminary univariate screening selected ~39 significant molecular features with *P*-value < 0.05 for the following steps. The proposed methodology outperformed other methods on all performance metrics and the FSS has the smallest expected model size except for the PCR that has super genes. The inclusion frequencies of molecular features for the molecular model are produced after the DCV procedure of STMC as shown in (a) of Figure 3.7. The single final model was build and assessed by LOOCV measures and KM curves are in (b) of Figure 3.7. The final model includes the five molecular features of Septin 1, thyroxine-binding globulin precursor,

KIAA0084 protein, cyclin-dependent kinase 7, and eukaryotic translation initiation factor2. The *P*-values of the log-rank test and the PI slope are very significant with P-values < <0.00001 and LOOCV *C*-index is 0.688.



(a) (b)

Figure 3.7 (a) Feature Relevance Ranking (FRR) for the molecular model of DLBCL data on mortality event, (b) Final model assessment for DLBCL data.

3.5 Simulation study

3.5.1 Simulated data

We simulated genomic data with P=1000 features and N=240 subjects to

demonstrate the effectiveness of the proposed approach for building a predictive model that is optimized to have a smaller signature in the N << P problem settings. All predictor values are generated from a uniform distribution [0, 1]. Tthe prognostic index of a linear risk score function $f(\mathbf{x})$ is formed for $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^{\mathrm{T}}$ so that only the true predictors are related to risk time to events. The survival time T is generated from an exponential distribution with parameter $\exp(f(\mathbf{x}))$ when $\mathbf{x}$ is given, and the censoring variable C is generated from an exponential distribution with parameter 0.4. Then we obtain the survival data, $\{(t_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i)) | i = 1, \ldots, n\}$ with right censoring effect and the EPV of simulated data is 0.84.

Table 3.7 Comparative performance analysis of methods for simulated data (85 variables; *P*-value < 0.05).

	Dataset	Simulated Data				
	Method	FSS	LASSO	Ridge	PCR	Proposed
Simulated Model	Exp. Model size	21(0)	36.3(7.92)	85(1.43)	*3.3*(1.34)	11.3(4.40)
	C-index	0.857(0.06)	0.861(0.06)	0.872(0.06)	0.862(0.09)	**0.893**(0.08)
	IAUC	0.903(0.05)	0.910(0.05)	0.911(0.04)	0.890(0.08)	**0.931**(0.10)
	IBS	0.200(0.10)	0.190(0.16)	0.269(0.15)	0.186(0.06)	**0.167**(0.07)
	DD	9.32(29.42)	-0.53(7.42)	-10.14(2.36)	-3.67(5.79)	**-13.54**(3.23)

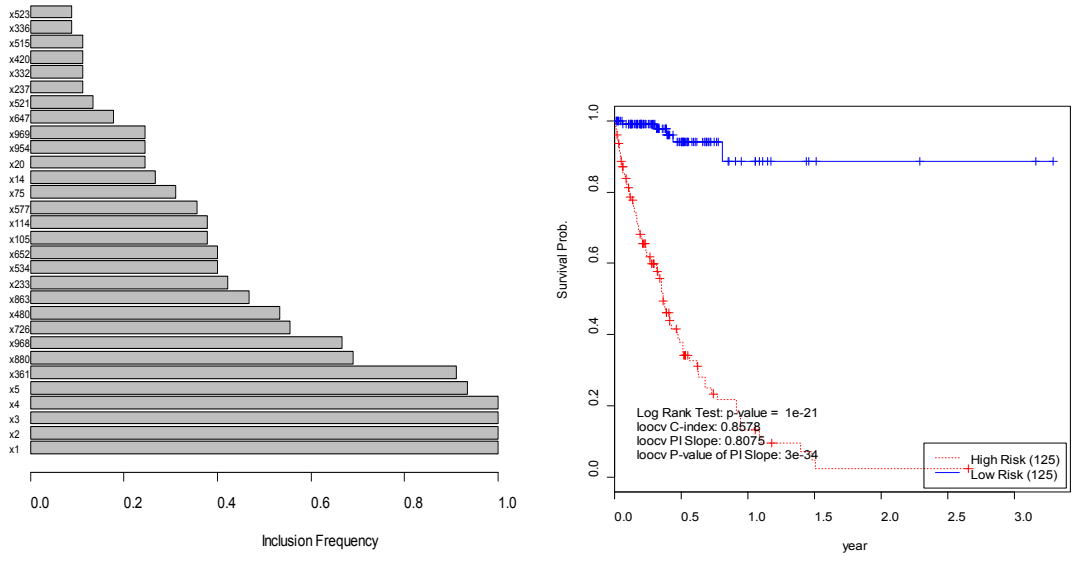
- The smallest expected model size is indicated in *italics*.
- Method performances which are the best are indicated in bold.

3.5.2 Experimental results

The prognostic models for the simulated data were constructed in the identical way with the above. Table 3.7 shows the comparative performance analysis of methods for the

simulated data. On average, 85 significant features with *P*-value < 0.05 were selected by the preliminary univariate screening for the following steps. The proposed approach also outperformed other methods on all performance measures and the expected model size of PCR was the smallest but was in lack of interpretability.

The inclusion frequencies of simulated features are produced after the DCV procedure of STMC as shown in (a) of Figure 3.8. The single final model was build and assessed by LOOCV measures and KM curves are in (b) of Figure 3.8. The final model includes the true five simulated features, x_1-x_5, and the other features, x_{361}, x_{480}, and x_{726} that are irrelevant but significant. However, both *P*-values of the log-rank test and the PI slope are very significant with P-values < <0.00001 and LOOCV *C*-index is 0.858.



(a) (b)

Figure 3.8 (a) Feature Relevance Ranking (FRR) for the simulated data, (b) Final model assessment for the simulated data

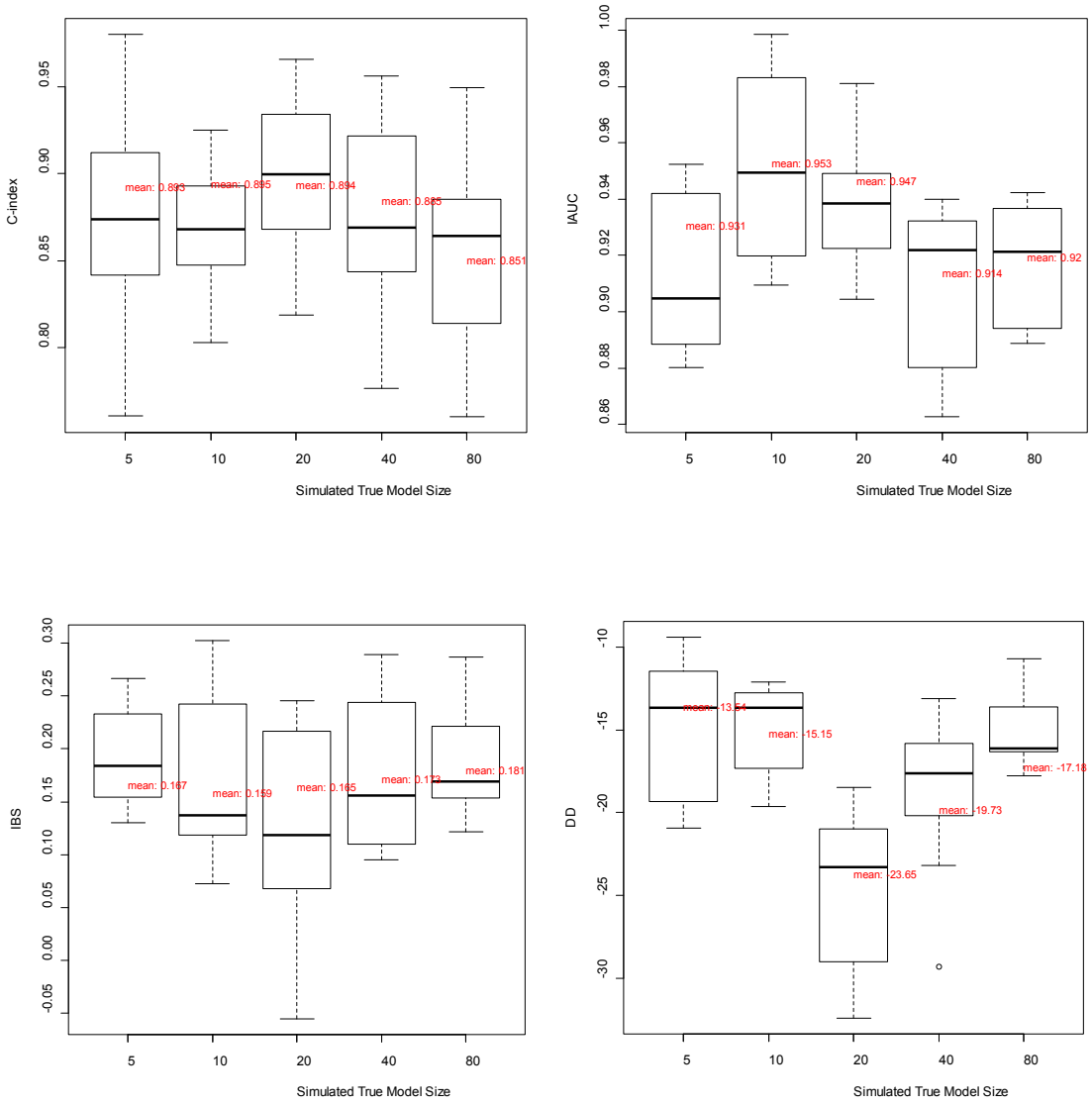


Figure 3.9 The performance analysis of our approach for five simulated datasets with the variation of the true model size of 5, 10, 20, 40, and 80 on the *C*-index. IAUC, IBS, and DD.

In addition to the above simulated data, we also simulated four additional genomic datasets of P=1000 features and N=240 subjects, whose the true model sizes are 10, 20, 40, and 80. Figure 3.9 shows the performance analysis of our approach for five simulated datasets with the variation of the true model size of 5, 10, 20, 40, and 80 on the *C*-index, IAUC, IBS, and DD. We analyze the results based on the mean value of each measure instead of using the median value. The true model size of 10 in simulation datasets was the best on the *C*-index, IAUC, and IBS except for the DD in which the size of 20 was the best. The performance of each measure tends to increase or decrease toward making worse as the size is augmented due to the multicollinearity among features.

3.5 Disscussion

In recent years, the rapid increase in availability of high dimensional gene expression profiles and the active research in translational science have given rise to the need for a data integration methodology to build a prognostic model to improve predictive accuracy. However, a primary problem of overfitting arises from the P >>N problem and censoring effect when obtaining estimates using the standard Cox model. We have presented a methodology using problem-oriented strategies to build an accurate prognostic prediction model with parsimony, using the data integration scheme that combines clinical and molecular features, which is reliable and is well calibrated for future prediction based on a trial application using highly correlated breast cancer data.

In order to avoid *overfitting* due to high dimensional features and small sample size settings, we employed the dimension reduction methods including the preliminary

univariate screening using a permutation test and QR decomposition. In particular, the correction for *overoptimism* in the situation of the P >>N problem is critical in experimental design if comparison with other methodologies is to be fair. For this reason, we have proposed a modified DCV for internal validation for comparison of methods, and have used replicated K-fold CV to find the optimal model complexity of a hybrid signature and to estimate the optimism corrected model performance. Predictive models were evaluated by several measures based on discrimination, calibration, and the overall score to reflect and consider the variability in the *censoring effect* when comparing models. Also, gene expression measurements in microarray data are highly correlated with one another. Thus, we used MPPLLE for L2 penalization in order to avoid *multicollinearity*, which adds estimation bias through coefficient shrinkage but reduces variance and, consequently, improves predictive accuracy as a result. Regression modeling for time to event data is much more sensitive to the event per variable than to the overall sample size and researchers carefully guide the 10 events per variable ratio [67]. As this is, however, a problem we cannot avoid when P >>N, we attempted to use the DCV and dimension reduction techniques to diminish estimation bias while simultaneously reducing the number of features. We have proposed using the strategies above to build a single final model with a hybrid signature as an optimal subset. For practical use of the developed model, the entire breast cancer data set must fit the hybrid signature.

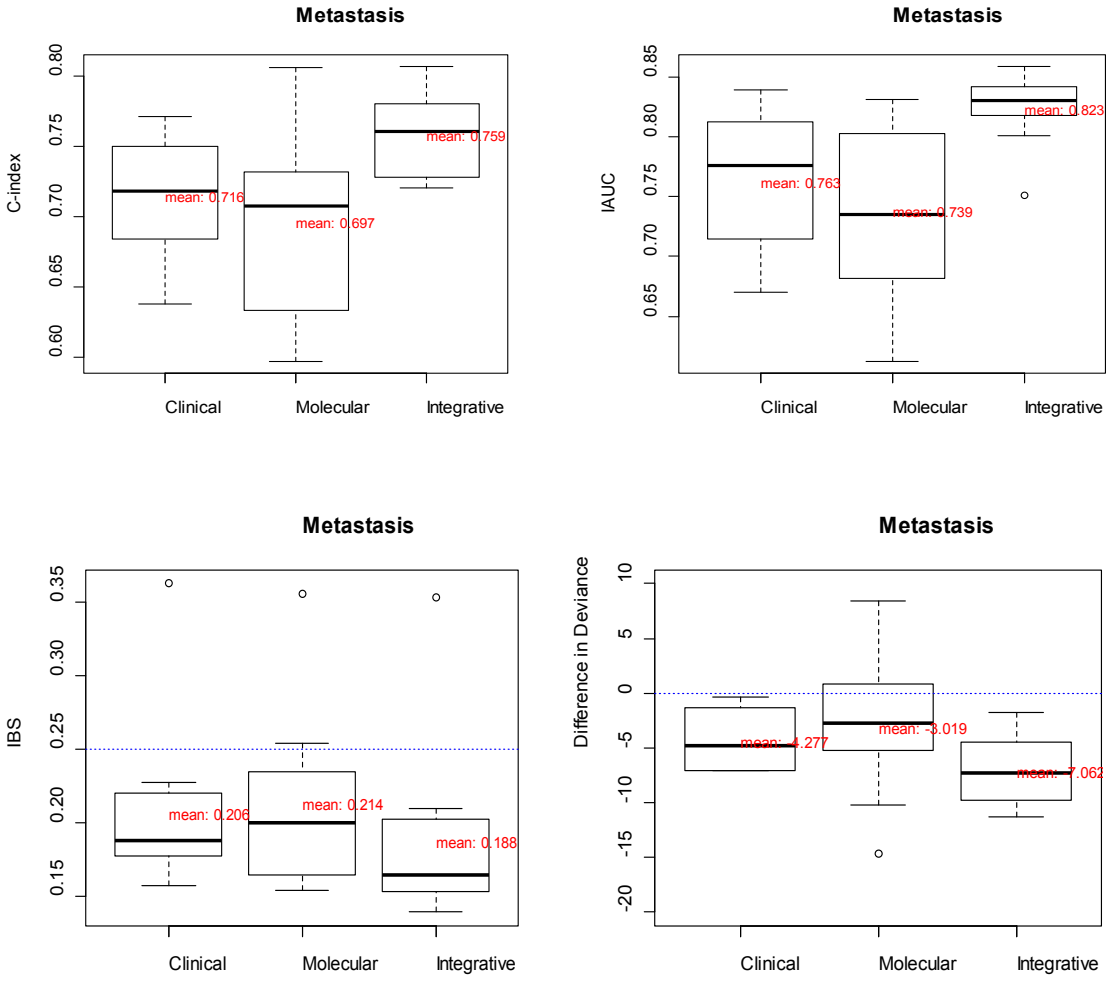


Figure 3.10 Performance analysis of data integration for the proposed method on metastasis outcome in breast cancer data.

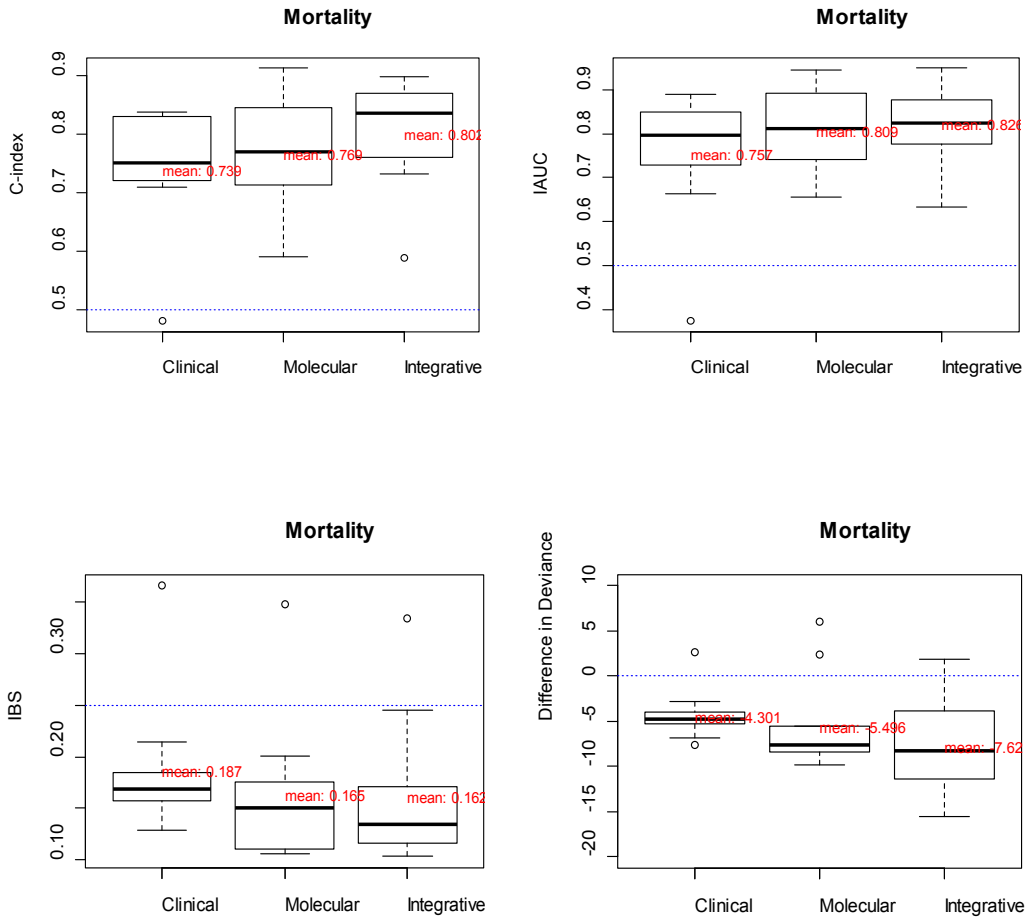


Figure 3.11 Performance analysis of data integration for the proposed method on mortality outcome in breast cancer data.

Our comparative study is based on the prediction of metastasis and mortality in the breast cancer data set. A comparative analysis for model building methods was performed using the FSS, LASSO, ridge, and proposed method. Our results on both events exhibited that our approach outperformed the other methods in the molecular and integrative

models, and was better on most of the measures than the other methods in the clinical model. Although the FSS size except the PCR, whose results include supergenes, was the smallest for the majority of the models, its performance deteriorated as the size decreased. The proposed method had a sufficiently small number of features so as to improve the performance. We should note that for metastasis and mortality outcome, the FSS method showed the poorest performance on the DD in the molecular and integrative model and those results imply that molecular features were very highly correlated and were not handled by penalization techniques.

Figure 3.10 and Figure 3.11 show the box plots of performance analysis of our data integration scheme for metastasis and mortality outcome and illustrate that the clinical model was better than the molecular model for metastasis but it was on the contrary for mortality, and overall, the integrative model outperformed them on all of the measures. In particular, the results for the metastasis outcome illustrate the stability of the integrative model with the smaller variance of most measures.

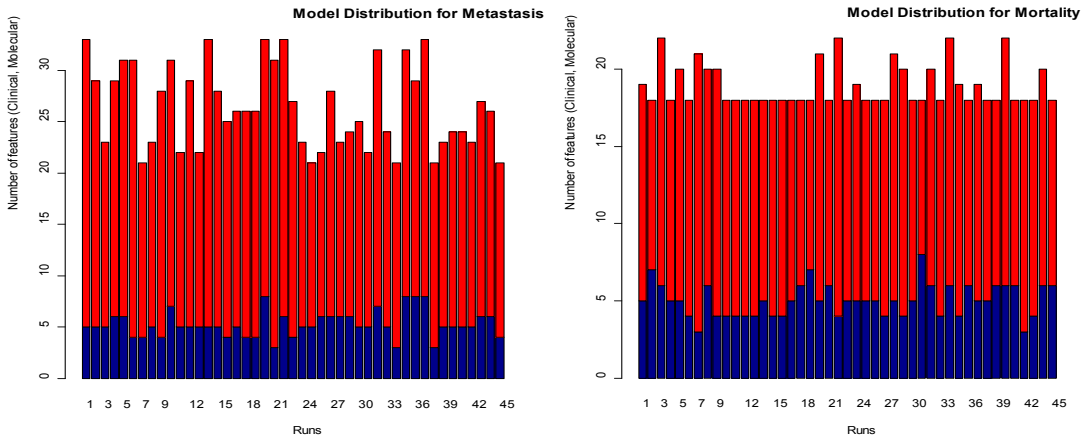


Figure 3.12 The analysis of selected clinical (blue) and molecular (red) features in the integrative model for metastasis and mortality on breast cancer data.

Figure 3.12 displays the number of clinical and molecular features selected in the integrative model for metastasis and mortality in each run of 45 models built in our methodology and shows that the feature numbers are stable in either type. In the single final model assessment, it was not easy to determine whether the clinical or molecular model on either event outcome performed better, but it was evident that the integration of clinical data and high-dimensional genomic data yielded better predictive performance than when each data set is handled separately, which was demonstrated by the KM curves and several LOOCV measures. Also, the experimental results on the DLBCL data and simulation studies reinforced that our methodology performed better than other methods.

For internal validation, we have used the DCV and a resampling method, the CV, which are relative but effective to achieve predictive performance for method comparison. For an accurate and independent performance evaluation, we may need additional

publicly available censored survival data for external validation and may need to perform bootstrap 0.632+ estimator to obtain accurate performance measures [50]. However, this bootstrap approach can demand a large number of resampling samples.

CHAPTER 4 Conclusions

In this dissertation, first, we proposed a new approach to construct accurate final clinical models with the optimally reduced size of risk factors through validation using resampling-based techniques in the high events per variable setting. The proposed method, STMC, compared to the stepwise selection methods with the different criteria of the LRT and the AIC, and LASSO, demonstrated better results in the two different data sets and a simulation study, and can be used for clinical prognostic modeling. The final model of FNSS improved the *C*-index at least better than the full model and had better performance on most measures. Also, in order to handle the high dimensionality and small sample size problem in censored time to event data, we have presented and tested an integrative model building methodology that improves the prognosis prediction accuracy for breast cancer data, and that optimizes to identify a small subset of a hybrid signature of clinical factors and molecular biomarkers that is most relevant to the risk of two clinical outcomes. When comparing methods using the DCV procedure, our proposed approach outperformed the other competing methods and the ridge Cox regression, which currently is accepted as having the best prediction performance, in the molecular and integrative models using breast cancer data. The results of the method comparison and the final model assessment for our methodology indicate that the discrimination between performance of clinical and molecular models depends on a dataset with a risk event, and that the integrative model improved predictive accuracy

over models that used clinical or molecular data alone. Also the unidentified genes among the features we selected can be investigated further to validate biological processes in breast cancer studies.

Although our approach is computationally demanding, we can obtained an integrative clinicogenomic model that is improved by predictive accuracy and includes a reduced and reliable hybrid signature in contrast with other methods.

Furthermore, the experimental results of the DLBCL data and simulated data supported our demonstration. Finally, although our approach of this dissertation utilizes a Cox model as a regression problem in censored time to event data, it can also be applied to other classification problems and binary outcome data using appropriate performance measures. This is because our methodology is based on a wrapper approach.

Appendix A

Description of Original Renal Transplantation Data (20085 records and 67 variables)

	Variable Names	Description	Type	Values(Min vs Max or % or #)	# of NA
1	TRANSPLANT_ID	ID Number of Kidney Transplantation	numeric(numeric)	6-231543	0
2	TX_DATE	Transplant Date	factor(numeric)	01/01/2000-12/31/2003	0
3	FAILDATE_KI	Date of Graft Failure	factor(numeric)	01/01/2001-12/31/2004	18680(93%)
4	PX_STAT_DATE	Date of Death, Re-TX or Last Follow-Up	factor(numeric)	01/01/2002-12/31/2005	1
5	wt	Weight of Recipient at Tx (*kg*)	numeric(numeric)	18.70-166.90	910(5%)
6	ht	Height of Recipient at Tx (*cm*)	numeric(numeric)	99.06-225.00	910(5%)
7	AGE	Recipient Age in years (yrs)	numeric(numeric)	18-84	0
8	bsa	Recipient Body Surface Area (m^2)	numeric(numeric)	0.7632-2.1250	0
9	bmi	Recipient Body Mass Index	numeric(numeric)	15.04-49.92	0
10	ht_inch	Recipient Height (inch)	numeric(numeric)	39.00-88.58	0
11	wt_lbs	Recipient Weight at Tx (lbs)	numeric(numeric)	41.14-367.20	0
12	racecat	Recipient Race	factor(numeric)	Black(2992:15%), White(13525:67%), or Other(3568:18%)	0
13	GENDER	Recipient Gender	factor(numeric)	F(8320:41%) or M(11765:59)	0
14	wtht	Recipient Preop Weight/Height Ratio	numeric(numeric)	0.1764-0.8940	910(5%)
15	diab_r	Recipient Diabetes	numeric(binary)	0(99.8%) or 1(0.2%)	5917(29%)
16	dial_reg	On Dialysis at Registration	numeric(binary)	0(33%) or 1(67%)	509(3%)
17	dial_tx	On Dialysis at Transplant	numeric(binary)	0(26%) or 1(74%)	322(2%)
18	fail_acu	Graft Failure Contrib. Cause: Acute Rej	numeric(binary)	0(91%) or 1(9%)	18864(94%)
19	fail_chr	Graft Failure Contrib. Cause: Chron Rej	numeric(binary)	0(91%) or 1(9%)	19167(95%)
20	gfail	Graft Failure	numeric(binary)	0(17785:89%) or 1(2300:11%)	0
21	dead	Recipient Death	numeric(binary)	0(18963:94%) or 1(1122:6%)	0
22	iv_dead	Time Interval (yrs) to Death or Censoring	numeric(numeric)	0-5.621	1
23	d_wtht	Donor's Weight/Height Ratio	numeric(numeric)	0.2516-0.8431	3239(16%)
24	d_wt	Donor's Weight (*kg*)	numeric(numeric)	39-161	3239(16%)
25	d_ht	Donor's Height (*cm*)	numeric(numeric)	106-221	3239(16%)
26	AGE_DON	Donor's age (yrs)	numeric(numeric)	15-48	0
27	GENDER_DON	Donor's Gender	factor(numeric)	F:11806 or M:8279	0
28	d_bsa	Donor's bsa	numeric(numeric)	1.293-2.978	0
29	d_bmi	Donor's bmi	numeric(numeric)	15.21-49.95	0
30	d_htinch	Donor's Height (inch)	numeric(numeric)	41.73-87.01	0
31	d_wt_lbs	Donor's Weight (lbs)	numeric(numeric)	85.8-354.3	0
32	d_creat	Donor Serum Creatinine (SCr) - (Preop)	numeric(numeric)	0.2-18.4	0
33	d_procec	Donor Procedure: Nephrectomy Type	factor(numeric)	Laparoscopi(13057:65%) or Open(7028:35%)	0
34	dracecat	Donor Race	factor(numeric)	Black(2767:14%), White(13876:69%), or Other(3442:17%)	0
35	abo_ordc	ABO Match	factor(numeric)	Compatible(4568:23%) or Identical(15517:77%)	0
36	gen_m2m	Gender: Male to Male Transplant	numeric(binary)	0(77%) or 1(23%)	0
37	gen_f2f	Gender: Female to Female Transplant	numeric(binary)	0(77%) or 1(23%)	0
38	gen_m2f	Gender: Male to Female Transplant	numeric(binary)	0(82%) or 1(18%)	0
39	gen_f2m	Gender: Female to Male Transplant	numeric(binary)	0(65%) or 1(35%)	0
40	iv_opyrs	Interval (yrs): 01/01/2000 to Tx	numeric(numeric)	0-3.997	0
41	iv_wait	Time (yrs) on Waiting List	numeric(numeric)	0-12.41	7795(39%)
42	im_thera	Induction Therapy-Depleting & Receptor	numeric(binary)	0(8919:45%) or 1(10912:55%)	254(1%)
43	im_fk506	fk506 (Tacrolimus) Maintenance	numeric(binary)	0(8776:44%) or 1(11055:56%)	254(1%)
44	im_calco	Calcineurin Inhibitor without fk506	numeric(binary)	0(12270:62%) or 1(7561:38%)	254(1%)
45	crcl_pr	CrCl (Cockcroft-Gault Formula) (Preop)	numeric(numeric)	0.6987-186	1588(8%)
46	gfr_pr	eGFR (MDRD) Pre-Transplant	numeric(numeric)	1.577-167.1	1271(6%)
47	crc_po6	CrCl (Cockcroft-Gault) in 12 Mths (Post-Tx)	numeric(numeric)	0.8795-748	7153(36%)
48	crc_po12	CrCl (Cockcroft-Gault) in 6 Mths (Post-Tx)	numeric(numeric)	0.9328-760.5	8368(42%)
49	kidrand	Random # Seeded 15789473 on 05/24/06	numeric(numeric)	0.0000823-0.9999	0
50	HLAMIS	HLA Mismatch Level	numeric(numeric)	0(11%),1(6%),2(19%),3(29%),4(13%),5(15%),6(8%)	0
51	dial_1wk	Dialysis in the First Week (Post-Tx)	numeric(binary)	0(95%)-1(5%)	0
52	trt_rej6	Any Treated for Rejection within 1st 6 mths	factor(numeric)	Y(2650:13%) or N(17435:87%)	0
53	diaggrpc	Primary Diagnosis: Cause of Renal Failure	factor(numeric)	Diabetes:3095,glomerulonephritis:4854,other:12000,retransplant:136	0
54	im_deple	Induction with depleting antibodies	numeric(binary)	0(16354:81%) or 1(3731:19%)	0
55	im_il2	Induction with IL2 Receptor Antidodies	numeric(binary)	0(12481:62%) or 1(7604:38%)	0
56	im_myco	Mycophenolate Mofetil Maintenance	numeric(binary)	0(4670:23%) or 1(15415:77%)	0
57	im_rapa	Rapamycin (Sirolimus) Maintenance	numeric(binary)	0(17125:85%) or 1(2960:15%)	0
58	im_aza	Azathioprine Maintenance	numeric(binary)	0(19362:96%) or 1(723:4%)	0
59	im_calci	Calcineurin Inhibitor with fk506	numeric(binary)	0(1356:7%) or 1(18729:93%)	0
60	gfr_po6	eGFR (MDRD) in 6 mths Post-Tx	numeric(numeric)	1.508-550.2	0
61	iv_gfail	Time (yrs) to Graft Failure or Censoring	numeric(numeric)	0-5.574	0
62	gfaildea	Graft Failure or Death	factor(numeric)	Dead(1122:6%),Gfail_alive(1178:6%),No_Gfail_alive(17785:88%)	0
63	gfr_po12	eGFR (MDRD) in 12 Mths Post-Tx	numeric(numeric)	3.225-988.7	0
64	PRAMR	RH Most Recent PRA (%) Pre	numeric(numeric)	0-100	0
65	PRAPK	RH Peak PRA (%) Pre	numeric(numeric)	0-100	0
66	drelgrpc	Relation of Donor to Recipient	factor(numeric)	biol_blood_related:13912,nonbiol_related:3657,spouse_partner:2516	0
67	iv_dial	Duration of Dialysis Pre-Tx (yrs)	numeric(numeric)	0-68.44	0

- BSA: Body Surface Area
- BMI: Body Mass Index
- eGFR: Estimated Glomerular Filtration Rate
- CrCl: Creatinine Clearance
- HLA: Human Lymphocyte Antigen
- PRA: Panel Reactive Antibody
- SCr: Serum Creatinine

Bibliography

[1] Ambler, G. and Brady, A.R. and Royston, P. (2002) Simplifying a prognostic model: a simulation study based on clinical data. *Statistics in Medicine*, 21(24):3803-3822.

[2] Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006) Prediction by supervised principal components. *J Am Stat Assoc*., 101, 119-137.

[3] Bellman, R. E. (1961) Adaptive Control Processes. Princeton University Press.

[4] Blum, A. and Langley, P. (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245-271.

[5] Bøvelstad, H.M., Nygård, S, Størvold, H.L., Aldrin, M., Bolgan, Ø., Frigessi, A. and Lingærde, O.C. (2007) Predicting survival from microarray data – a comparative study. *Bioinformatics*, 23:2080-2087.

[6] Bøvelstad, H.M., Nygård, S, Bolgan, Ø. (2009) Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics*, 10(413).

[7] Cecka JM. (2005) The OPTN/UNOS Renal Transplant Registry. *Clinical Transplants* ,1-5.

[8] Chang, HY., Nuyten, DSA., Sneddon, JB., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He, YD., van't Veer. LJ., Bartelink, H., Rijn, M., Brown, PO., Vijver, MJ. (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci USA*, **102**(10), 3738-43.

[9] Choi, I., Wells, BJ., Yu, C., and Kattan, MW. (2011) An empirical approach to model selection through validation for censored survival data. *Journal of Biomedical Informatics*, in press.

[10] Collett, D. (2003) Modeling Survival Data in Medical Research. Chapman and Hall, second edition.

[11] Cox, D.R. (1972) Regression models and life tables (with discussion). *J.R. Stat. Soc. B*, 34, 187-220.

[12] Daemen, A., Gevaert, O., and De Moor, B. (2007) Integration of clinical and microarray data with kernel methods, *Medicine and Biology Society*, 5411-5415.

[13] Gene H. Golub , Charles F. Van Loan (1996) Matrix computations (3rd ed.), Johns Hopkins University Press, Baltimore, MD.

[14] Gevaert, O. et al. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinfomratics*, 22, 184-190.

[15] Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K., and van Houwelingen, J. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9), 1950-1957.

[16] Goeman, J. J. (2010) L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1):70-84.

[17] Goldblatt, EM., Lee, WH. (2010) From bench to bedside: the growing use of translational research in cancer medicine. *Am J Transl Res*, 2(1), 1-18.

[18] Gonen, M., and Heller, G. (2005) Concordance probability and discriminatory

power in proportional hazards regression. *Biometrika*, 92(4):965-970.

[19] Graf E, Schmoor C, Sauerbrei W, and Schumacher M. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*,18:2529-2545.

[20] Guyon, I. and Elisseeff, A. (2004) An introduction to variable and feature selection. *JMLR*, **3**, 1157–1182.

[21] Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Bontempi, G. (2008) A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*, 24(19):2200-2208.

[22] Harrell, F. E., Lee, K. L., Mark, D. B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.

[23] Harrell, FE. (2001) Regression Modeling Strategies, Springer-Verlag.

[24] Hastie, T., Tibshirani, R., and Friedman, J. (2001) The elements of statistical learning. Springer, New York.

[25] Heagerty, P. J., Lumley, T., and Pepe, M. S. (2005) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56, 337-344.

[26] Heath, M. T. Scientific computing: an introductory survey. (2002) McGraw-Hill, Boston, MA, second edtion.

[27] Hielscher, T., Zucknick, M., Werft, W., and Benner, A. (2010) On the prognostic

value of survival models with application to gene expression signatures. *Statistics in Medicine*, 29(7-8):818-29.

[28] Kattan, MW., Wheeler, TM., Scardino, PT. (1999) Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J Clin Oncol*, 17(5), 1499-1507.

[29] Kattan, M. W., and Gonen, M. (2008) The prediction philosophy in statistics. *Urologic Oncology*, 26(3):316-319.

[30] Kattan, MW., Vickers, AJ., Yu, C., Bianco, FJ., Cronin, AM., Eastham, JA., Klein, EA,. Reuther, AM., Pontes, JE., Scardino, PT. (2009) Preoperative and postoperative nomograms incorporating surgeon experience for localized prostate cancer. *Cancer*, 115(5), 1005-1010.

[31] Kohavi, R. and John, G. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273-324.

[32] Li, L. (2006) Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22, 466–471.

[33] Mann, H. B., and Whitney, D. R. (1947) On a Test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*,18(1):50-60.

[34] Matsui, S. (2006) Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinformatics*, **7**(156).

[35] *McNeil. (1982) The meaning and use of the area under a receiver operating*

characteristic (ROC) curve, radiology, 143:29-36.

[36] Miller, A. (2002) Subset Selection in Regression. Chapman and Hall, second edition.

[37] Molinaro, A.M., Simon, R., and Pfeiffer, R. M. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301-3307.

[38] Nguyen, D. V. and Rocke, D. M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*. 18, 1625-1632.

[39] Nygård, S. *et al.* (2006) Partial least squares Cox regression on genomic data handling additional covariates. *Statistical Research Report* 5/2006. Department of Mathematics, University of Oslo.

[40] Ohno-Machado L. (2001) Modeling medical prognosis: survival analysis techniques. *J Biomed Inform*, 34(6):428-39.

[41] Park, M.P., and Hastie, T. (2006) L1 regularization path algorithm for generalized linear models. *Technical report*. 2006-14. Department of Statistics, Stanford Univeristy.

[42] Peduzzi, P., Concato, J., Feinstein, A. R., et al. (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*, 48, 1503-10.

[43] Porzelius, C., Schumacher, M., and Binder, H. (2010) A general, prediction error-based criterion for selecting model complexity for high-dimensional survival

models. *Statistics in Medicine*, 29(7-8):830-838.

[44] Raykar, V., Steck, H., Krishnapuram, B., et al. (2007) On ranking in survival analysis: bounds on the concordance index. *In Advances in Neural Information Processing Systems*, 20:1209–1216.

[45] R Development Core Team. R. (2008) A Language and Environment for Statistical Computing, Vienna, Austria.

[46] Rosenwald A, Wright G, Chan WC, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, 346:1937-1947.

[47] Saeys, Y., Inza, I., Larrañaga, P., (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23, 2507-2517.

[48] Sauerbrei W. (1999) The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics*, 48:313-329.

[49] Sauerbrei, W, and Royston, P. (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society*; Series A, 162: 71-94.

[50] Schumacher, M., Binder, H., Gerds, T., (2007) Assessment of survival prediction models based on microarray data, *Bioinformatics*, 23, 1768–1774.

[51] Stephenson, AJ., Smith, A., Kattan, MW., *et al.(2005)* Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, 104, 290-298.

[52] Steyerberg, E. W., Eijkemans, M. J., Habbema, J. D. (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol*, 52(10):935-942.

[53] Steyerberg, E. W., Eijkemans, M. J., and Van Houwelingen J. C., et al. (2000) Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine,*19:141–60.

[54] Steyerberg E. W., Eijkemans, M. J., Harrell, F. E., Jr, Habbema, J. D. (2000) Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine,* 19(8):1059-79.

[55] Steyerberg, E. W., Harrell, F. E., and Borsboom, G. J., et al. (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*, 54 :774–81.

[56] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen M., Obuchowski, N., Pencina, M.J., Kattan, M.W. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1), 128-138.

[57] Sun, Y., Goodison S., Li, J., Liu, L. and Farmerie, W. (2007) Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23, 30–37.

[58] Therneau, T. M. and Grambsch, P. M. (2000) Modeling Survival Data: Extending the Cox Model. Springer.

[59] Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385-395.

[60] Tiong, H. Y., Goldfarb, D. A., Kattan, M. W., Alster, J. M., Thuita, L., Yu, C., Wee, A., Poggio, E. D. (2009**)** Nomograms for predicting graft function and survival in living donor kidney transplantation based on the UNOS Registry. *J Urology*, 181(3):1248-55.

[61] *Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani R., Botstein, D., and Altman RB. (2001*) Missing value estimation methods for DNA microarrays. *Bioinformatics,* 17, *520-525*

[62] van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999-2009.

[63] Van Houwelingen, HC. (2000) Validation, calibration, revision and combination of prognostic survival models, *Statistics in Medicine*,19(24), 3401-3415.

[64] Van Houwelingen HC, Bruinsma T, Hart AA, Van't Veer LJ, Wessels LF. (2006) Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine,* **25** , 3201-3216.

[65] van't Veer, LJ., Dai, H., van de Vijver, MJ. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415, 530-536.

[66] Varma, S., and Simon, R. (2006) Bias in error estimation when using cross-

validation for model selection. *BMC Bioinformatics*, 7(1):91.

[67] Vittinghoff E.and McCulloch, CE. (2006) Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epi*, 165**,** 710-718.

[68] Wilcoxon, F. (1945) Individual comparisons by ranking method. *Biometrics. Bull.*, 1, 80-83.