Prognostic Factors and Predictions of Survival Data

by

Duo Zhou

A Dissertation Submitted to

The Rutgers University

School of Health Related Professions

In partial fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Department of Health Informatics

November 5, 2014

RUTGERS
BIOMEDICAL AND HEALTH SCIENCES

Final Dissertation Approval Form

Prognostic Factors and Prediction of Survival data

BY

Duo Zhou

Dissertation Committee:

Dinesh P. Mital, PhD, Committee Chair

Masayuki Shibata, Ph. D.

Shankar Srinivasan, PhD, Committee Member

Approved by the Dissertation Committee:

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

ABSTRACT

Survival outcome has been one of the major endpoints for clinical trials; it gives information on the probability of a time-to-event of interest. There has been increasing interest in survival analysis tools over the recent years, especially for high dimensional survival data. Common statistical approaches include nonparametric, semi-parametric and complete parametric analysis, several of which are widely used and readily available from major commercial software applications. However most of these approaches have limitations. Typical nonparametric approaches, such as the log-rank (or Cox-Mantel) test, are not concerned about model assumptions, but can only deal with a limited number of categorical predictors. Typical semi-parametric approaches, such as Cox proportional hazard model, depend very much on the model assumptions, such as linearity, interactions and proportionality; also these approaches can only deal with survival data when the number of predictors is less than the total number of events. Complete parametric models, such as accelerate failure time models, are similar to semi-parametric models except that they make further assumptions about the baseline hazard function.

In this research paper, we studied several techniques for evaluating survival data, the typical Cox PH models including the generalized Cox linear model and the multivariate Cox regression models with nonlinear transformations, the nonparametric random survival forest approaches, penalized Cox regression models including lasso, ridge and elastic-net Cox regression models, derived-input Cox regression models including principal component Cox regression and partial least squares Cox regression models. These models were implemented and evaluated with one simulation study and one real world case study.

The typical Cox models including the generalized Cox linear model and the multivariate Cox regression models with nonlinear transformations should always provide unbiased estimates, and the models are flexible for handling recurrent-event survival response; but they are incapable of making inferences for cases when there are more predictors than the actual number of events; and since they are semi-parametric

approaches, model assumptions such as linearity, interaction and proportionality, should be carefully examined before the models were implemented. In this paper, a systematic procedure was proposed for examining the model assumptions, which should help to ensure the correct model was employed for the survival data. In terms of prediction performance, they are among the best approaches.

In the paper, we also introduced nonparametric random survival forest approaches, log-rank based and conditional inference based random survival forest models, which have many advantages over the typical nonparametric, semi-parametric or parametric approaches. There are no concerns about model assumptions, and these methods can deal with many more predictors than typical survival models. In terms of prediction performance, these models are moderate and slightly worse than the typical Cox models.

The penalized Cox regression models, on the other hand, should always give biased estimates; but they work quite well for cases when the number of factors is no less than the number of events. Of all penalized Cox models, the elastic-net Cox model works extremely well for correlated high dimensional data; the prediction performance is extremely good. However, they do not work for multiple event type of survival data.

The principal component Cox regression model is a very useful tool for variable reduction with similar prediction performance as the typical Cox models. The model also has similar features as the typical Cox models; it can deal with recurrent event or interval censored survival data. But it also has many disadvantages, in cases when the number of components is no less than the total number of observations, the model may not be estimable; more importantly, analysis results from this model may be difficult to interpret.

The partial least squares Cox regression model was developed; it shares some resemblance with principal component Cox regression model, the only difference is the construction of the components, instead of the building orthogonal components independent from the survival outcome, the model builds the PLS components to attain the strongest correlation with the survival outcome, otherwise it has similar features as the principal component Cox regression model. Additionally, the prediction performance of this model is unexpectedly very disappointing.

ACKNOWLEDGEMENT

I would like to take immense pleasure in thanking to my advisor, Professor Dinesh P. Mital for his valuable guidance and advice; he convincingly conveyed a spirit of adventure in research and scholarship, and an excitement in teaching. Without his persistent help this dissertation would not have been possible.

I would like to thank to my department chair, Professor Syed S. Haque, who introduced me to health informatics, who also taught me how to cherish time. I would like to offer my sincere appreciation for the education opportunities and encouragement from him.

I would like to express sincere gratitude to my former manager, Eric Yan, for his abundant generosity to allow me continue my academic journey and who also offered invaluable advice on balancing my work, family and academic life.

I would like to gratefully acknowledge the inspiration from Professor Masayuki Shibata who taught me on how to utilize computer aided analysis for drug development; his support and encouragement had given me the sparkling idea for this research.

Finally, to my caring, loving and supportive wife, Liya, whose encouragement and support had got me to complete this dissertation. It was a great comfort and relief for me to complete my work while she was helping my entire family and managed the household activities.

TABLE OF CONTENTS

LISTING OF TABLES

LISTING OF FIGURES

List of Abbreviations

SBP	Systolic blood pressure
DBP	Diastolic blood pressure
BMI	Body mass index
MAP	Mean arterial pressure
RCS	Restricted cubic spline
AIC	Akaike information criterion
AUC	Area under the receiver operating characteristics curve
ROC	Receiver operating characteristics
KM	Kaplan-Meier
AFT	Accelerated failure time
SVT	Support vector machine
PH	Proportion hazard
PCA	Principal component analysis
BMI	Body mass index
df	Degree of Freedom
AG	Andersen-Gill
PWP	Prentice, Williams and Peterson
WLW	Wei, Lin and Weissfeld
CI	Confidence interval
iid	Independent and identically distributed
PSH	Peña, Strawderman and Hollander
AECM	Alternating expectation conditional maximization
EM	Expected-maximization
WC	Wang and Change
VIF	Variance inflation factor
PL	Product limit
HR	Hazard ratio
PLS	Partial least squares
LOESS	Locally weighted scatterplot smoothing
MTV	Maximum total variance
SOC	Standard of care
RGLM	Random generalized linear model
Bagging	Bootstrap aggregation
OOB	Out-of-Bag
IPCW	Inverse probability censoring weight
RSF	Random Survival Forest
VIMP	Variable importance
CV	Cross validation

Chapter 1. Introduction

1.1 Background

Over the past decade, a large amount of data has been collected across many disciplines. One of the most important components is the biomedical data, of which, the aggregation has undergone remarkable growth. The questions, how to extract useful information or evidence for better diagnosis or prediction from the existing data, how to discover hidden relationships or identify susceptible subpopulation from the existing data, and how to best utilize the existing data for future research and discovery, have led to the prosperity of knowledge discovery techniques such as data mining, machine learning and statistical learning.

Data can be distinguished by the quantitative or qualitative (categorical) feature and it can be further subdivided into continuous or discrete, nominal, ordinal data or survival data. While in medical research, the outcome of interest can be binary, continuous, ordinal, counting process and survival outcomes, of which time-to-event outcome has been one of major endpoints for clinical trials; it gives information on the probability of a time-to-event of interest.

Depending on the types of outcomes, different statistical analysis approaches must be intended, including approaches based on parametric, semi-parametric or nonparametric models. Typical nonparametric approaches are based on little or no model assumptions and can only deal with a limited number of categorical predictors. Typical parametric, semi-parametric and other approaches, such as statistical learning techniques based on parametric models are built under some model assumptions, including but not limited to the distribution assumptions for outcome of interest, the distribution assumptions for variables to be included in the model, underlying relationships between factors and outcomes. As a result of the parametric modeling, statistical inferences can be estimated intuitively from the analysis. Since these approaches are built with some assumptions, violation of the assumptions, may lead to questionable inferences[1, 2, 3] and unexpectedly large prediction errors. Therefore, when violation of the assumptions is detected, effort should be sought to transform the data in order to satisfy the assumptions or alleviate the violation.

However, there are times when data transformation cannot solve or reduce the violation, when alternative applications should be intended, such as some procedures to relax the parametric assumptions, some approaches using parametric or semi-parametric models without similar stringent assumptions or with different assumptions, or other methods based on nonparametric models without or with little assumptions. No matter what type of models, they have to be stable, dependable and reliable, i.e., the same model from different processes should achieve similar or equivalent results.

On the other hand, nonparametric approaches can be used across several different platforms; they are built with no or very little assumptions, such as recursive partition[4], random forest[5], boosting[6] and etc.; nevertheless, the approaches are limited for statistical inferences, and relatively more difficult to interpret.

1.2 Previous Publications

As for continuous and categorical outcome, extensive research has been carried out and the typical analysis approaches are quite efficient for prognostic factor detection and robust for prediction. In 2011, Zhou et al. [7] used a proportional odds logistic regression analysis to identify prognostic factors for open angle glaucoma and later in 2013, Zhou et al. [8] applied a statistical learning technique based on logistic discriminant models to predict the malignant breast cancer with extraordinary sensitivity and specificity.

Since then, the research interest has been focused on prognostics factor analysis and predictive modelling for survival data.

1.3 Focus of the Thesis

Survival data arises in many fields, such as medicine, biology, public health, epidemiology, engineering, economics and demography. The tools and approaches presented in research paper should be general and applicable to all of these disciplines, but the focus of the research was survival data from biology and medicine.

Frequently, survival outcomes are always collected when the intension is to study serious disease conditions, such as death, heart failures and recurrence of cancers. The occurrence of such conditions may be referred to as events or failures. Speaking of events of failures, some may take a long time to occur or may not occur within certain period of time, while others may occur within a short time; some may occur only once, while

others may occur multiple times; some events of the same type may occur repeatedly, while others may be accompanied by many other types of events. For these type of data, the outcome of interest typically includes a binary variable to indicate the occurrence of the conditions or events and a continuous variable to indicate the time of the occurrence(s) or the time of censoring (sometime, two continuous variables may be needed to indicate the interval within which the event or censoring occurs). These types of responses are also referred to as time-to-event outcomes.

Although, the survival outcome consists of two (or more) variables; the continuous variable, time of the occurrence, can be evaluated using a typical multivariate regression analysis, and the binary variable, the occurrence of the event, can be assessed with a typical logistic regression to model. Yet, it is insufficient to draw conclusions based on either one of the analyses and it is almost impossible to get an overall picture with consideration of the results from both analyses. To best utilize time-to-event data, completely different statistical models are built to exploit the complete information about the occurrence of the event and the time of the event occurrence, with the intention to determine the probability of event to occur within a specific time, the probability of recurrence and/or the gaps between the occurrences.

The intension of such models may be laid upon prediction of probability of the future occurrence within certain time, discovery of the contributing prognostic factors to the serious conditions, and/or identification of the subpopulation susceptible to the conditions of interest. The primary interest is to analyze survival data; the outcome includes the time-to-event from a certain cause, duration of response to treatment, time to disease recurrence, time to adverse event or simply time-to-death. As such, different survival models to bridge the statistical theory and medical research are developed, studied and compared to evaluate the prediction performance.

To build models for survival data, common features of survival data have to discussed, such as censoring and/or truncation, which have contributed to the complexity of the data. Furthermore, different schemes of censoring and/or truncation have made the modeling even more difficult. Different combinations of censoring and/or truncations schemes have to be modeled differently. Hence, it is essential to study the nature of the censoring and truncation schemes.

Censoring exists when an individual lives through the study or discontinues the study without experiencing the event of interest. Three types of scheme for censoring have to be considered for survival analysis, right censoring, left censoring and interval censoring.

Right censoring scheme occurs when a subject has left the study or study has ended before the event occurs. There are three subtypes for right censoring: type I, type II and random censoring. Type I censoring occurs when subjects enter the study at different time points and the study is followed for a pre-specified duration; the censoring time is the pre-specified "terminal point" for subjects who do not experience any event. Type II censoring occurs when the study ends as soon as a pre-specified number of events is collected; subjects may enter the study at different times, but they all terminate at the same time when the pre-specified number of events has occurred; the censoring time for subjects without any events is determined by the time when they enter the study. Another type, random censoring or progressive censoring, may occur when other competing events have caused subjects to be removed from the study randomly. As can be seen, type I and II censoring cannot occur at the same time, but either may occur with random censoring. For example: in a study with pre-specified follow-up time, subjects may not have had events before they dropped out the study due to unwillingness to participate, the random censoring occurs when the subject drops out before he even has any chance to have an event, for the rest of the subjects who do not experience any events, the type I censoring occurs at the time when they discontinue at the end of the pre-specified follow-up. Right censoring is the most commonly reported censoring scheme in clinical studies and medical research.

Left censoring occurs when a subject has experienced the event of interest prior to the start of the study, but the exact time of the event is unknown. An example of this scheme was reported by Turnbull and Weiss (1978)[9], a study was conducted in California to determine time to the first marijuana use among high school boys; in the study, a question was asked "when did you first use marijuana?" One of the responses was "I have used it but cannot recall just when the first time was." A boy who chose this response indicated that the event had occurred prior to the interview but the exact time at which he started using marijuana is unknown. Left censoring scheme is not common in clinical studies, it is relatively more common in survey studies and usually accompanies with right

censoring scheme. In the above example, if one of the responses is "I never used it", then it is the right censored observations at the boys' current age.

Interval censoring scheme arises when the exact time of the event cannot be obtained but can only be determined to occur within an interval from examination performed on scheduled follow-up visits; this is another commonly reported schema in clinical studies. Furthermore, right censoring is just a special case of interval censoring, when the left and right bounds of the intervals are equal to each other, i.e. both bounds of the interval are set to the exact time of event occurrence.

Truncation is a variant of censoring where subjects are included in a study only if they survive until the start of the study or if events have occurred by a given date; truncation is usually caused by a systematic selection process from the study design. There are two schemes for truncation, right and left truncations. Right truncation occurs when a subject has experienced the event of interest before study entry. Left truncation arises when the subjects have survived or have been at risk for a sufficient time before entering the study.

Studies to evaluate time-to-event data are usually conducted within a pre-specified follow-up duration; theoretically, if all subjects are followed long enough (without any limitation), everyone should experience the event of interest sooner or later. A practical survival model has to consider that the event of interest may not have occurred within the pre-specified follow-up period for some subjects; in other words, the model has to account for cases that the occurrence of the event may occur after the follow-up ends (or subject drops out).

Therefore, when modeling survival data, at least two variables have to be considered, a binary variable to indicate the occurrence of the event of interest and a continuous variable to indicate the time of the occurrence of the event or the censoring time (sometimes, 2 continuous variables may be needed to indicate the interval within which the event occurs or the subject censors); if an event of interest has occurred before the follow-up ends or subject drops out, then it is flagged as a failure; otherwise if the event of interest has not occurred before the subject dropout or follow-up ends, it is flagged as censoring at the time of dropout or at the end of the follow-up. And the corresponding time is called event time and censoring time, respectively.

When the primary interest is to characterize the subgroup of subjects who are more likely to relapse after a surgery or who are prone to the side effect within certain period of time after treatment, prognostic factor analysis will play an important role. When the primary interest is to predict the probability whether the event will eventually occur, without concern too much about the time of the occurrence, then a pure classification model to predict the probability of event occurrence with the consideration of censoring shall be enough.

Therefore, survival models are usually more complex as comparing to statistical models for other data types; the methodology utilized for survival models is more specific in both modeling and the type of data. Without losing generality, it is assumed throughout the research paper that all subjects are independent of each other, only events occurring within the same subject are correlated, the studies are right-censored and censoring is independent of the event.

1.4 Evaluation Performance

Models may underfit as well as overfit the data. Sometimes if a model does not perform well for the training set and larger than expected error is observed, then it is considered as underfitting and the corresponding error is referred to as the bias of the model. Some other times, a model may fit the training set quite well, but it may not be able to accurately predict future events, which is considered as overfitting; it may occur when too much variation, random errors or noises, have been built into the model, therefore it is also referred to as variance. Overfitting generally occurs when a model is excessively complex and it may exaggerate normal fluctuations in the data. This is due to different criteria used for analysis and prediction. For analysis, a model is typically selected by maximizing the cross validation performance on the collected dataset, while prediction is evaluated based on unseen data. Practically, it is more convenient to train the model, evaluate the cross validation performance and examine the prediction performance without waiting for the unseen data, therefore a random procedure to divide the collected data into the training set (or cross validation set) and the test set should be employed.

The training set will be used to build and train the model; the test set will be used to examine the prediction performance of the mode. Usually cross validation can be

performed over the training set, for detection of underfitting; but sometimes a separate cross validation set may be needed for tuning the model (such as pruning a decision tree). Testing is the only and necessary set for detection of overfitting.

As mentioned previously in section 1.3, survival data are usually collected for evaluation of serious disease conditions; and the goal of every survival analysis is to generalize the model from the training example to all possible input. Considering the price for collection such data, it is always wise to have most if not all information for training, especially when the size of the training set may be crucial for detection of prognostic factors. Then a bootstrap procedure may be needed for testing or validation; the bootstrap will randomly sample the original survival data with replacement for testing and validation.

1.5 Objectives

For survival analysis, nonparametric models can only be used to adjust for a few categorical factors, thus they have limited usage for prognostic factor detection and prediction of future events. On the other hand, the semi-parametric Cox regression model has been very flexible and robust, even if a completely parametric model may be more suitable. However, it is noticed that this model has been frequently misused in medical research. Frequently, even if not always, factors are only included in the Cox regression model in their first order linear forms, no interactions or only interactions between the linear form of predictors are included, and time-dependent covariates are not considered, adjusted or addressed. Such simplified strategies are not completely wrong, they are certainly not perfect; statistical inferences or predictions based on such model will certainly be questionable or vulnerable. Sometimes, the number of predictors in the survival data is no less than the total number of events, the typical Cox regression model will not work. Furthermore, complete parametric models, such as accelerated failure time models are similar to the semi-parametric models except that they make further assumptions about the baseline hazard function. Therefore, how to properly model survival data for better detection of prognostic factors and how to make more accurate or reliable inferences and predictions, have puzzled researchers for years.

The objective of the research paper is to develop systematic approaches for appropriately modeling survival data, accurate detection of prognostic factors and robust

prediction of survival outcomes in order to bridge the statistical analysis with actual clinical practice. In this research, only right centered and left truncated data were considered, since these two schemes are most typical in clinical research.

For survival data, only the subjects who have experienced the event of interest will contribute to the analysis, the subjects who do not have any event will make little or no contributions to the analysis. However, even with a large datasets, there will always be a portion of subjects who may actually experience event(s). For a typical survival model, in order make reasonable estimates and ensure a proper fit of the model, the number of events will have to be at least 10 times more than the number of factors to be fit to the model without considering interactions and non-linear effect. Is there a systematic approach to utilize the survival data more efficiently without losing too much generality? Yet, combining with machine learning techniques, we were able to build models including more factors, interaction and non-linear forms; we managed to analyze survival data even if the number of factors was no less than the total number of event (see section 4.2 for the real world case study on NKI70 data).

Before any analysis on prognostic factors, for typical parametric survival models, there are always concerns about multicollinearity, heteroscedasticity, interactions, confounding factors and time-dependent or time-varying effect, which may have significant impact on statistical inferences and prognostic factors detection, even though they may or may not affect the model predictions. Therefore, are there any systematic diagnostic tools that we can employ to diagnose, reduce or avoid these problems without affecting the model performance? Are there any survival models that can properly address these issues with reasonable model performance?

For prognostic factors, are all factors always having linear relationships with the survival outcomes? Are all interactions between factors always linear? How to detect the nonlinear relationships? Additionally, some factors may have changed over the course of the disease, how to accommodate the changes for these factors and best utilize such information in model predictions.

Machine learning is a new branch for carrying out data analysis without worrying too much about statistics and/or assumptions in order to make reliable inference and predictions. For many machine learning techniques, they are not built on strong statistical

foundations, many statistical rules or principals may not be applicable.

With respect to survival models, most parametric statistical models are developed with some underlying assumptions, so that reasonable inference can be obtained from the analysis; some other approaches are based on nonparametric models with little or no assumptions, but inferences cannot be easily obtained. Are there systematical approaches to force the variables to satisfy the model assumptions or is there a way to relax the model assumptions so that we can still use the approaches based on parametric models to obtain reasonable inferences and predictions? Are there any approaches based on nonparametric models with little or no stringent assumptions for making predictions of future observations and perform formal analysis with reasonable performance? Are there any other approaches based on semi-parametric models with different or relaxed model assumptions for better inference and predictions?

For certain disease conditions, subjects may experience multiple events while others may only experience one event; while multiple events occurred within the same subject may be correlated with each other, or sometimes competing events may occur within the same subject. Are there any models to adjust for the within-subject correlations? When competing events are observed, some subjects may not experience any event before they are terminated by a single serious state, such as death. Are there any approaches for modeling survival outcomes while accounting for competing events?

For all intended models, how to evaluate the model performance, how to properly compare the performance of difference model, including the typical statistical approaches and the approaches based on machine learning techniques and how to interpret the results? It is a common belief that the more factors selected, the more accurate prediction can be achieved, is it really true? Are all factors making similar contributions to the survival outcomes, or are there any factors more important than others for prediction of the survival outcome?

1.6 Hypotheses

Statistics include many constituents other than hypothesis testing, such as study design, estimation, prediction and etc. The approaches discussed in this paper aims at develop systematic approaches for prognostic factors analysis on time-to-event data and/or accurate prediction of future survival outcomes; though some components may be

considered s superset of hypothesis testing and estimation, but others may be too complicated to be formulated as hypothesis testing, such as model performance and/or prediction of future outcomes, intra-subject correlations for multiple observations, diagnosis of nonlinearity, interactions and/or proportionality. Additionally, evaluating multiple hypotheses prior to application of formal analyses will end up spending extra degree of freedoms, which may further inflate the total variance and lead to overfitting; thus the ordinary point estimates arise from the hypotheses generating assessments are significantly biased because of "data over-dredging". Theoretically, such hypotheses generation processes are supposed to be avoided in modern statistics; however, when such steps cannot be avoided.

However statistics inferences and predictions are typically based on statistical models, some complicated task can be simplified using hypothesis tests to ensure proper fit. The following hypotheses are intrinsic within model building and model selection, which will assist in building the right and appropriate models for prognostic factor detection and predictions. Moreover, there are cases when statistical models involve too many predictors, hypothesis testing can be utilized as an alternative but convenient tool for automating the screening process.

- Normality Assumption of Predictors:

  For parametric or semi-parametric survival, normality is the basic assumption for all generalized regression analysis. If data are highly skewed, extra caution should be taken and data may need to be transformed. So normality check is the first step for most parametric or semi-parametric approaches.

  H_0: Predictor satisfies normality assumption;
  H_a: Normality assumption is violated.

  For normality, D's Agostino's K-squared test[10], Anderson-Darling test[11], Kolmogorov-Smirov test[12] and Shapiro-Wilk Test[13] may be performed.

- Nonlinearity:

  Do all predictors have linear relationship with log hazard? Does any factor have a nonlinear form in the survival model?

  H_0: Factor has a linear relationship with the hazard function;
  H_a: Nonlinearity exists.

Nonlinearity can be checked using hypothesis test by checking the p-value from the deviance difference between the fitted models with and without the nonlinear form. However hypothesis test for determination of functional form may lead to incorrect conclusion if a wrong form was pre-specified. For example, a quadratic form of the factors may not be detected with hypothesis testing; yet it can be easily detected with the graphic display of the Martingale residuals[14] against the linear form of the factor. Although hypothesis test may not be perfect, still it is a convenient tool when it is impossible or difficult to check the graphic display for every single factor.

- Interactions between Factors for Cox PH Mode:

  Are there any interactions between factors?

  H_0: There is not interaction between factors A and B;

  H_a: The interaction between factor A and B exists.

  Interactions can be evaluated using the p-value from the Wald tests by tentatively fitting to the Cox PH model, or it can be tested using the deviance difference between the fitted models with and without the interactions. Yet, both are not very robust, but they are convenient, which allow for systematical checks of the interactions using computer programs; otherwise, hypothesis tests as well as residual plot (scaled Martingale residuals) against each continuous factor stratified by the categorical variable may be more efficient.

- Proportionality Assumption for Cox PH Mode:

  Do all factors satisfy the proportionality assumption for the intended survival model?

  H_0: All predictors satisfy the proportionality assumption;

  H_a: At least one predictor does not satisfy the proportionality assumption.

  Proportionality assumption can be checked using a global Chi-square test[15] of the scaled Schoenfeld residuals[14] on a function of time, yet it is not very stable, instead a residual plot vs. time is much more powerful (a non-zero slope is an indication of a violation of the proportional hazard assumption).

- Multicollinearity Detection:

  Is there any multicollinearity among factors?

  H_0: $r_{ij} = 0$;

  H_a: $r_{ij} \neq 0$ for at least one pair of (i, j), $i = 1 \dots p, j = i \dots p$, and $i \neq j$.

Please note that p is referring the total number of factors (see section 2.1 for details). Correlation coefficient is just one measurement for multicollinearity, and there are other measurements, such as variance inflation factors. But detection of multicollinearity is only the first step, the next step is how to best utilize the information for modeling survival data.

- Nonparametric test on survival curves

  The survival curves from all groups are equal to each other; in other words, are there any differences in survival probability among different groups;

  H_0: All survival curves are the same.

  H_a: At least one survival curve is not the same as the others.

  If there are only 2 groups, the hypothesis test can be tested using log-rank test; if there are > 2 groups, Cox-Mantel l[26, 27] test must be applied (see section 2.3 for details).

- Goodness of Fit:

  Does the model fit; i.e., does the specified model describe the survival data appropriately?

  H_0: Model fits (or the model correctly describes the data);

  H_a: Model does not fit (or the model does not describe the data well).

  Goodness-of-fit is a general measurement for evaluating how well the model can describe the survival data; for Cox model, it includes several components as discussed previously, proportionality assumption, linearity, interactions and finding of the overly influential data point.

- Prognostic Factor Detection and/or Model Selection

  None of the covariates are significant for prediction of time-to-event outcome (probability of event within a certain period of time) vs. one or more covariates are significant for model prediction. In terms of each factor, is it a risk factor for predicting failure events of interest within certain period of time?

  H_0: $\beta_1 = \beta_2 = \cdots = \beta_\nu = 0$;

  H_a: $\beta_1 \neq 0$, or $\beta_2 \neq 0, \ldots,$ or $\beta_\nu \neq 0$ ($\beta_i \neq 0$ for at least one i, $i = 1, 2, \ldots, \nu$).

  Please note that ν (defined in section 2.1) is referring the total number of covariates, not the total number of factors.

A global Null hypothesis on $\beta = 0$ can be checked with three tests, global likelihood ratio test, Score test and Wald test; of the 3 tests, the global likelihood ratio test is the most reliable, then Score test and Wald test is the least.

Hypothesis test on β_i for covariate i can be incorporated into the model selection procedure to select contributing prognostic factors (this is one of the options for model selection, but it has to be admitted that the model selection procedure based on *P*-values violates statistical principals).

⋮

In addition to the above, some tools may be more powerful than hypothesis testing and some measurements just cannot be formulated as hypothesis tests, such as the detection of linear or nonlinear relationship (see section 3.8.3 for details), then visual inspection of the smooth spline plots should be more intuitive; for proportionality assumption and nonlinearity, graphical display of data can reveal hidden relationships that cannot be detected from hypothesis testing. Moreover, for model selection, it is always advisable to use adjusted Akaike information criterion (AIC) and adjusted R^2 for model selection than hypothesis tests; for model validation and calibration or performance, hypothesis tests are not available, alternative measurements should be evaluated (see section 3.3.2 for details).

However, hypothesis tests are still useful, when alternative tools are unavailable or difficult to assembly; i.e., if there are too many predictors, it is difficult to check the pair-wise interactions with graphic plots, but hypothesis tests can be implemented easily.

1.7 Research Significance

As mentioned previously, survival data are collected to assess the most serious medical conditions, such as death, heart failures, adverse side effects after initial treatment and recurrence of cancer cells after initial surgery or chemotherapy. In consideration of the seriousness of the conditions, researchers are more concerned about early diagnosis, so that preventative treatment can be administered before the condition is developed or deteriorated; for health care professionals, identify the susceptible subpopulation will assist in prescribing early treatment before the serious conditions become irreversible or futile; for scientists, prediction of the probability of time-to-event can help to prioritize and to develop personalized treatment; for insurance companies,

pre-screening and pre-detection of the serious conditions may help to save money by allowance of preventative treatment before the serious condition occurs. With this motivation, several analysis approaches for survival outcome were extensively studied in this research.

For survival data, the primary outcome of interest includes the status of the conditions, indicating the event occurrence and a duration variable (or variables) indicating the time (or the interval) of the event occurrence; additionally, survival models also have to account for cases when individuals drop out prematurely or do not have any event before the end of the study; moreover, there are many censoring and truncation schemes, combination of different schemes should make survival analyses even more complex. For analysis of survival data, all information has to be considered in the modelling, which makes survival outcome much more difficult to model than typical continuous or categorical outcomes.

Moreover, all statistical inferences are based on statistical models; incorrect models may lead to incorrect or biased inferences. Thus, the initial survival model is the basis for all successful statistical analysis; any further analysis procedures are descendent from the initial model. For example, model selection, prognostic factors detection and model prediction of further events are all derived from the initial survival model. However, it is not trivial to build an appropriate model without much knowledge about the model itself, especially when there are too many factors to consider.

The parametric accelerate failure time (AFT) models depend on stringent model assumptions, thus these models are not widely used. Similarly, the semi-parametric Cox models also have stringent model assumptions, they do not assume baseline hazard as the parametric AFT models, but they do make an additional assumption on proportionality.

However, due to the deficiency of effective tools, the model assumptions for parametric AFT or semi-parametric Cox models, are not always checked, therefore most of the analysis models may not be formulated appropriately; the inferences obtained from the inappropriate model may lead to incorrect conclusions. Furthermore, these models may sometimes run into non-estimability dilemma due to large number of parameters and small number of events. The typical nonparametric approaches however, have no or little assumptions, but they can only adjust for a limited number of categorical factors and

continuous factors have to be categorized before use in the analysis. Thus, all of these typical approaches have limitations.

With the above considerations, alternative approaches with the help of machine learning techniques, such as lasso[16, 17], ridge[98] and principal component Cox regression models have been proposed, but these approaches also have limitations, lasso Cox regression model is only useful when the number of parameters is no more than the number of observations; ridge regression does not select factors, therefore it is not very useful for prognostic factor detection; principal component Cox regression model relies on latent components which are constructed independent of the survival outcomes, therefore it is difficult to interpret the results with respect to the original factor, moreover it is not guarantee to be estimable.

In this research paper, we developed a systematic process to assess the assumptions for typical Cox regression models. During the process to assess the linearity assumption, appropriate functional forms were recommended if nonlinearity was detected; when assessing interaction effect, appropriate interactions were suggested if interaction effect was identified; when assessing proportionality assumption, appropriate time-dependent extension was integrated if non-proportionality was detected.

In addition, to overcome the disadvantages and the limitations of the semi-parametric approach (Cox model), we introduced nonparametric random survival forest to select the best subset of prognostic factors with moderate prediction performance; we also applied an innovative semi-parametric elastic-net Cox model for prognostic factor detections with excellent prediction performance; in addition, we also developed a partial least squares Cox regression model for highly correlated survival data, which could be used to adjust for uncollected covariates (latent components). All of these approaches are capable of dealing with the cases when the number of parameters is no less than the total number of observations.

Chapter 2. Literature Review

This chapter will review the definitions, formulations and statistical models for survival analysis including parametric, semi-parametric and nonparametric models. Section 2.2 provides an overview of the terms and definitions for survival analysis, including definitions of hazard function, cumulative hazard function, probability density function, cumulative distribution function and the survival function. Section 2.3 reviews commonly used survival analysis models, including nonparametric survival analysis, parametric survival models and semi-parametric survival analysis models. The nonparametric survival analysis include Kaplan-Meier (KM) estimator or product limit (PL) estimator, Nelson-Aalen estimator for survival probability or survival curves, log-rank test and Cox Mantel test for group comparison, Wang-Chang model for recurrent event survival analysis. The parametric analysis models, include exponential, Weibull, Gompertz and generalized Gamma, log-normal accelerated failure time (AFT) and log-logistic AFT survival models. The semi-parametric survival analysis models include Cox proportional hazard (PH) model, and several extensions or variations of Cox PH model. Section 2.4 reviews the generalized regression analysis and section 2.5 reviews several survival models based on machine learning techniques.

2.1 Naming Conventions

Several conventions are used throughout the paper, only those terms that are not straightforward and have special hidden indication will be emphasized in this section. For example, factors or predictors will be referring to the actual variables or features collected from the survival data; the total number of factors is denoted by p. Covariates or terms will be referring to the actual terms included in the survival model; i.e., for linear form, each of the original factors will be used as the covariate term for the survival model; for nonlinear transformations, each term of the nonlinear forms of the factors will be used as a covariate term for the survival model; interaction terms will be constructed between the original factors or the nonlinear transformation of the factors as well as time-dependent effect (option 2 from section 3.8.5); the total number of covariate terms is denoted by v.

Random variables are denoted as T or X capitalized. The capitalized letter T is used to

denote the random survival time; t is used to denote the vector of time for each subject for analysis of time-to-event outcome; the capitalized letter X denote the matrix of covariates, with each covariate from different subjects as a column vector and with different covariates from the same subject as a row factor. For typical Cox proportional hazard model, the input matrix does not include the column vector of 1's, since the Cox regression model is based on partial likelyhood, it assumes no baseline hazard function. Thus, the Cox PH regression model does not have an intercept (see section 2.3.3.1 for details). However, the baseline hazard function may be specified for parametric proportional hazard models; i.e., the input matrix, X, should include a column vector of 1's as its first column corresponding to the intercept for regression analysis for parametric exponential proportional hazard models. Lower letter, i, is usually used for index of time when an event occurs; lower letter, j, is usually used for index of subject; lower letter, l, is usually used for index of clusters; except otherwise as noted.

Greek letter, β or α is used to denote the regression coefficient vector for covariate matrix, with each entry in the vector as a parameter coefficient for each covariate (column) in the design matrix, X. $Y(T|X)$ is a generic notation for the transformed response, which is the link function for the regression analysis; it can be a vector for typical survival analysis, or matrix for multiple outcomes (such as recurrent and terminal events). For Weibull model, $Y(T|X) \equiv \log[-\log(S(t|X)]$, $\log[\lambda(t|X)]$ or $\log[\Lambda(T|X)]$; for log-logistic model, this is $Y(T|X) \equiv \log\{S(t|X)/[1 - S(t|X)]\}$; for log-normal model, $Y(T|X) \equiv \log(t|X)$; for Cox proportional hazard, $Y(T|X) \equiv \log \mathrm{HR}$. The symbol, $\equiv$, means equivalent. When used in lower letters, $y(t|X)$ and x, will be used as scalar function (or vector function for time-to-recurrence and terminal events) or a linear combination of all covariates for a particular subject, respectively.

2.2 Overview of Terms and Definitions for Survival Data

Survival data are used to measure time-to-event of interest, such as failures, deaths, occurrences of a given condition, event recurrence (relapse), occurrences (or recurrences) of competing events. The time variable(s) are subject to random variations, and like any other random variables, form a distribution. The distribution of survival time is usually described or characterized by one of the 5 functions: (1) survival function, (2) density

probability function, (3) cumulative probability function, (4) hazard function and (5) cumulative hazard function. These five functions are mathematically equivalent —if one is given, the other four functions can be automatically derived. More importantly, they form the basis for all survival models. In practice, the five functions can be used to illustrate different aspects of the survival data. A basic problem in survival analysis is to obtain estimates of one or more of these five functions from the collected sample of survival data and to draw inferences about the survival pattern in the population[18].

Above all, the formulations of the 5 functions are derived from the definition of the entities; no underlying distributions are assumed. Hence, they are still considered as nonparametric and can be applied across all survival models.

The survival function $S(t)$ is defined as the probability of survival beyond time, t;

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(x)dx = 1 - F(t) \quad \text{..... Eq. 1}$$

The event density function, $f(t)$

$$f(t) = F'(t) = \frac{d}{dt}F(t) = \lambda(t)\, exp[-\Lambda(t)] \quad \text{..... Eq. 2}$$

The lifetime (cumulative) distribution function, $F(t)$

$$F(t) = Pr(T \leq t) = 1 - S(t) \quad \text{..... Eq. 3}$$

The hazard function, $\lambda(t)$, is the probability of event (event rate) at time t for subjects at risk prior to that time conditional on survival beyond time, t.

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt \mid T \geq t)}{S(t)dt} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \quad \text{..... Eq. 4}$$

The cumulative hazard function, $\Lambda(t)$:

$$\Lambda(t) = \int_0^t \lambda(x)\, dx = -\log S(t) \quad \text{..... Eq. 5}$$

The survival function can also be expressed as the inverse function of the cumulative hazard function, $\Lambda(t)$

$$S(t) = \exp[-\Lambda(t)] \quad \text{.....Eq. 6}$$

Other than the 5 basic entities, the following statistics can also be estimated if the distribution of survival time is known; the mean survival time μ can be estimated using Eq. 7 and the corresponding 100(1-α)% CI can be estimated using the variance of the means using the standard variance formula.

$$\hat{\mu}(T) = \int_0^T \hat{S}(t)\, dt \quad \text{.....Eq. 7}$$

$$\hat{\text{var}}(T) = \int_0^T t^2 dt - \mu^2$$

The qth sample percentile of survival time and the corresponding 100(1-α)% CI can be obtained as (The median survival time, $T_{0.5}$, can be obtained with $q = 0.5$):

$$T_q = \frac{1}{2}\left(\inf\{t: 1 - \hat{S}(t) \geq q\} + \sup\{t: 1 - \hat{S}(t) \leq q\}\right) \quad \text{.....Eq. 8}$$

$$I_q = \{t: -z_{1-\frac{\alpha}{2}} \leq \frac{\hat{S}(t) - (1-q)}{\sqrt{\text{var}\left(\hat{S}(t)\right)}} \leq z_{1-\frac{\alpha}{2}}\}$$

2.3 Review of Survival Models

Of all survival analysis approaches, the product-limit (PL) method, developed by Kaplan and Meier in 1958, has been one of the oldest nonparametric approaches for estimating survival function and it is still broadly used today. Another similar nonparametric approach, Nelson-Aalen estimator, was introduced for estimation of cumulative hazard function by Nelson (1972). For comparing two survival curves, log-rank test was introduced in 1966 and Cox-Mantel test was later developed to compare more than 2 survival curves. Not until the last decade, Wang et al. and Pena et al. introduced an extension of the product limit estimator for studying recurrent event survival data. Even though more parametric models have been developed today, the above nonparametric approaches are still broadly used, not only for estimation of survival function, but for inspection the distribution of survival time as well. Just recently, several of these nonparametric models have been implemented using machine learning algorithms to detect prognostic factors for survival outcomes.

Other than the nonparametric approaches, several parametric models may be used for analysis of survival data, such as exponential, Weibull, log-normal and log-logistic models. Except for exponential model, for which the hazard rate is a constant over time, the rest of the parametric models allow hazard to change over time with assumptions of probability distributions for the survival time, this is why they are also referred to as accelerated failure time (AFT) model. These models are as efficient and robust as Cox PH model if the data satisfy all parametric assumptions.

Nevertheless, the semi-parametric, Cox proportional hazard (PH) model is probably still the most popular approach for survival analysis. The imminent progress in machine learning algorithm has triggered the development of diversified approaches for survival analysis, most of them are built on top of Cox PH model.

Moreover, many extensions of Cox PH model have been proposed for broader applications in survival analysis, such as time-dependency, recurrent events, competing events models etc.

2.3.1 Nonparametric Estimation of Survival Curves

Nonparametric or distribution-free methods are quite easy to apply; comparing to parametric models, they are less efficient if the survival time follows a theoretical distribution and they are more efficient if the underlying distributions for survival time is unknown. Therefore, nonparametric models are often used as alternative tools to compliment the parametric or semi-parametric approaches. Especially for identification of the underlying distribution of the survival time, estimates and plots of the survival probabilities as well as the cumulative hazards from the nonparametric analysis should be very helpful.

2.3.1.1 Kaplan-Meier (KM) and Nelson-Aalen Estimator

Edward L Kaplan and Paul Meier (1958) were among the first to develop the Kaplan-Meier (KM) estimator[19] for estimation the survival probability for time-to-event data; it is also known as product limit estimator, which is a step function with steps at the event times when the event actually occurs. The estimation process takes advantage of the counting process, stochastic integral at each observed time (at the censoring time, the count of events will be 0; the count will be ≥1 only at the event times). Even though this is a simple estimation process, it is still widely used for preparing survival curves as a convenient nonparametric approach to estimate the survival probability beyond time, t; the actual formation for estimation of survival probability is presented in Eq. 8. The cumulative hazard function can be estimated accordingly base on Eq. 5 (section 2.2). For estimating the confidence intervals (CI) for KM estimator, several different formulas were proposed. The most popular one was proposed by Greenwood[20, 21]; another version, "exponential" Greenwood formula was proposed by Hosmer et al.[21] in 1999.

Both formulas will be discussed later in the section.

The Kaplan-Meier method can be used to estimate the survival probability or probability of event for the observed time period without any assumption of the underlying distribution. This is also the simplest concept for estimating the survival probability when there are no covariates. Although simple, it forms a platform for understanding the more complex models and theories. The probability of surviving at time t_k can be estimated using the product of the k observed survival rates for each $t_i, where\ i = 1 \dots k$. Let p_i denote the proportion surviving the period $[t_{i-1}, t_i)$ with $t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_k$, then $p_i = \frac{r_i - d_i}{r_i}$, where r_i is the number at risk at t_{i-1}. Then the survival probability at t_k, can be estimated by

KM Estimator	$\hat{S}(t_k) = \prod_{i=1}^{k} p_i$ ..... Eq. 9

The likelihood function can be formulated as

$$L = \prod_{i=1}^{k} (1 - p_i)^{d_i} p_i^{r_i - d_i} \quad \text{..... Eq. 10}$$

CI for survival probability at time t_k can be estimated as $\hat{S}(t_k) \pm \mathrm{z}_{\alpha/2}\sqrt{\mathrm{var}\left[\hat{S}(t_k)\right]}$, where the variance can be estimated by Greenwood's formula

$$\mathrm{var}\left[\hat{S}(t_k)\right] = \left[\hat{S}(t_k)\right]^2 \sum_{i=1}^{k} \frac{d_i}{r_i(r_i - d_i)} \quad \text{.....Eq. 11}$$

A second approach (exponential Greenwood formula) for estimation of CIs is obtained through log negative log of the survival probability as $\log\{-\log[\hat{S}(t_k)]\} \pm \mathrm{z}_{\alpha/2}\sqrt{\hat{V}_k}$, where $\hat{V}_k$ is the variance of the log negative log of the survival probability at time t_k

$$\hat{V}_k = \frac{1}{\{\log[\hat{S}(t_k)]\}^2} \sum_{i=1}^{k} \frac{d_i}{r_i(r_i - d_i)} \quad \text{.....Eq. 12}$$

The second approach (exponential Greenwood formula) is more preferred over the traditional Greenwood CIs, as it guarantees the upper and lower bounds of the CI to lie within (0, 1). However, in finite samples, the exponential Greenwood formula may not be

a good in the tails either. In particular, if the last subject in the sample has an event, the estimator and the variance are infinite.

An alternative nonparametric estimator, Nelson-Aalen estimator[22], was introduced for estimation of cumulative hazard function by Nelson in 1972, this estimate also utilized the stochastic integral at the observed times. In terms of survival probability, the Nelson-Aalen estimator and Kaplan-Meier estimator are asymptotically equivalent; but when sample size is small, Nelson-Aalen estimator performs better. On the other hand, the Nelson-Aalen estimator is commonly used to check assumptions for parametric models and get crude estimates for hazard function.

The formulation of Nelson-Aalen estimator of the cumulative hazard function is displayed below and the estimator of survival probability can be further derived from Eq. 6 (section 2.2; the corresponding 100(1-α) CI of the cumulative hazard function can be estimated asymptotically using the formula, $\hat{\Lambda}(t_k) \pm z_{\alpha/2}\sqrt{\mathrm{var}[\hat{\Lambda}(t_k)]}$.

Nelson-Aalen Estimator

$$\hat{\Lambda}(t_k) = \sum_{i=1}^{k} \frac{d_i}{r_i} \quad \text{..... Eq. 13}$$

$$\mathrm{var}[\hat{\Lambda}(t_k)] = \sum_{i=1}^{k} \frac{d_i}{r_i^2}$$

$$\mathrm{var}\left[\hat{S}(t_k)\right] = \left[\hat{S}(t_k)\right]^2 \sum_{i=1}^{k} \frac{d_i}{r_i^2}$$

2.3.1.2 Log-rank Test and Cox-Mantel Test (Nonparametric)

Once the survival probabilities are obtained, difference between survival probabilities is the next to estimate. For parametric and semi-parametric approaches, the difference is frequently expressed as hazard ratio, which will be discussed later in section 2.3.2 and 2.3.3. For nonparametric approaches, the difference is often assessed using hypothesis test against equal survival curves. In 1966, Nathen Mantel first introduced log-rank test[23, 24, 25] for assessing the differences between survival curves, which has become one of the most popular tools for comparing two survival functions. Another similar test, Log-rank form of Cox-Mantel[26, 27] test is capable for comparing difference among treatment groups of ≥ 2; there are several options for calculating the Cox-Mantel test statistics.

When there are only 2 groups, usually both tests give nearly identical results. But the results can differ when there are ties (events from multiple subjects occur at the same time), as pointed out by Bernsetin et al. in 1981[28] in a simulation study, neither method is accurate. They found that log-rank test tends to report equivalence (accept the null hypothesis as listed in section 1.6) and Cox-Mantel test tends to claim significance (reject null hypothesis as listed in section 1.6).

The log-rank test[23, 24, 25] is used for comparison of two survival curves. The χ^2 test statistics for log-rank test can be obtained as

$$\chi^2 = \frac{\left[\sum_{i=1}^{k}(d_i^{(a)} - r_i^{(a)} d_i/r_i)\right]^2}{\sum_i^k \frac{r_i^{(a)} r_i^{(b)} d_i (r_i - d_i)}{r_i^2 (r_i - 1)}} \quad \text{.....Eq. 14}$$

where the superscripts $(a)\ and\ (b)$ refer to one of the two groups; k is the number of distinct times that the events of interest were observed and i refers to the time i.

A more general form of this test (Cox-Mantel) statistics for comparison of the survival curves among u $(u \geq 2)$ groups $(j = 1, 2, \ldots u)$ is

$$\chi^2_{u-1} = \mathrm{U}' \mathrm{V}_w^{-1} \mathrm{U} \quad \text{.....Eq. 15}$$

where U can be estimated by $\mathrm{U}_i = \sum_{j=1}^{u} w_i (d_{ji} - e_{ji})$; $\hat{V}_w = w\hat{\mathrm{V}}w$, $(\hat{\mathrm{V}}_i)_{jj} = \frac{n_{ji}(n_i - n_{ji}) d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$ and the $(\hat{\mathrm{V}}_i)_{lh} = \frac{n_{li} n_{hi}\, d_i (n_i - d_i)}{n_i^2 (n_i - 1)}, h, l = 1, 2, \ldots u - 1$. The choice of weight, w_i, will lead to the choice of different options for Cox-Mantel test statistics.

The weight w_i for the log-rank (Peto et al.[29] 1972) test is 1. For generalized Wilcoxon test, the weight, w_i, is n_j following the Gehan-Breslow[30, 31] method, the square root of n_j following the Tarone-Ware[32] method and the Kaplan-Meier estimates of the survival probability multiplied by $n_j/(n_j + 1)$ following the Peto-Prentice[33, 34] method. The expectation of the number of events in group j at the ith time where d_i events is observed, can be estimated as, $e_{ji} = n_{ji} d_i / n_i$, where $n_{ji} = r_{ji} + d_{ji}$. Of the above methods, log-rank is more powerful if hazard rates are constant over time; the Gehan-Breslow Wilcoxon test is more powerful if the initial event rates (hazard rates) are higher, but it is less powerful if event rates are constant over time[35].

2.3.1.3 Nonparametric Wang-Chang Estimator of Recurrent Event

The nonparametric approaches discussed in the previous sections are all used to evaluate time-to-single-event data; however, in real world studies, multiple events survival data are very common in public health, epidemiology, medicine and clinical trials. Frequently, after multiple event survival data are collected, only the first occurrence of the events of interest from each experimental subject is used for most survival analysis (including the nonparametric analysis discussed above and most of the parametric and semi-parametric survival analysis). As a result, the extra information from the other events occurred in the same subject will be lost.

To model the recurrent event survival data, complex nonparametric models for recurrent events have been proposed with the only assumption that the distribution is smooth and multiple in both the time-to-event and covariates [36, 37].

In 2001, Peña et al.[38] proposed a nonparametric estimator as an extension to the product limit estimator (which was named as PSH estimator after the names of the authors) for recurrent event survival data with the only assumption that the inter-occurrence times are iid from some underlying distribution. This assumption was obviously very stringent for clinical research, thereafter the generalized frailty model was proposed to allow for association among inter-occurrence times.

For frailty distribution, a convenient choice is the gamma distribution with shape and scale parameter equal to an unknown parameter, α. Then, the marginal survival function can be written as, $F = [\alpha/(\alpha + \Lambda_0(t)]^{\alpha}$. The parameter, α, reflects the degree of association between the inter-occurrence times of the multiple events within the same subject. Peña et al. (2001) demonstrated that the estimation of α and $\Lambda_0(t)$ could be obtained via maximization of the marginal likelihood function using the expected-maximization (EM) algorithm. The inter-occurrence times are assumed to be from an independently and identically distributed (iid) sample following some underlying distribution. To obtain a better convergence, α was tentatively estimated; the initla estimates for α was used as a starting point in the EM procedure; the maximization of the profile likelihood for α was further estimated using a "golden section search method".

For the same reason, Wang and Chang (1999) proposed another nonparametric approach to obtain the estimator (this is named as Wang-Chang or WC estimator after the

names of the authors) of the common marginal survival function with adjustment of within-subject correlations among inter-occurrences. Wang and Chang claimed that this model should work with any frailty distributions, even though they considered an iid distribution with multiplicative properties. As such, gamma and other frailty models were just special cases.

When the inter-occurrence times are correlated within subject units, this model eliminates the bias of the estimates as noted for the product-limit estimator developed by Peña et al in 2001. However, for iid inter-occurrence times, this approach does not perform as efficient as the PSH estimator.

Compare the two nonparametric estimator, Peña, Strawderman and Hollander and Wang-Chang estimators (PSH and WC) for recurrent event survival data, WC estimator is preferably more reasonable in practice; however it only considers the situation that the initial occurrence of the event is an inclusion criterion for recruitment and the recurrences of the same type of events are observed within a pre-specified period of time. Therefore, the time of the initial occurrence of the event is defined as the origin of time (time=0).

Let i be the index for an event; j be the index for a subject; T_{ji} denote the time from $(i-1)$th event to the ith event for subject j, where $i = 1, 2, \dots, K_j$ and $j = 1,2, \dots, n$. Let C_j be the censoring time or the time between the initial event to the end of the follow-up for subject j; let G be the survival function of C_j and let $K_j = \{T_{ji}: i = 1, 2, \dots\}$. It is reasonable to assume that $(N_1, C_1), (N_2, C_2), \dots, (N_n, C_n)$ are iid. Let m_j denote the index satisfy $\sum_{i=1}^{m_j-1} T_{ji} \leq C_j$ and $\sum_{i=1}^{m_j} T_{ji} > C_j$. Let $m_j^* = \begin{cases} 1 & \text{if } m_j = 1 \\ m_j - 1 & \text{if } m_j \geq 2 \end{cases}$ and $U_j(t) = \frac{1}{m_j^*}\sum_{i=1}^{m_j^*} I(T_{ji} > t)$, then the mass of event set at time t is

$$d^*(t) = \sum_{j=1}^{n}\left[\frac{a_j I(m_j \geq 2)}{m_j^*}\sum_{i=1}^{m_j^*} I\,(T_{ji} = t)\right] \quad \text{..... Eq. 16}$$

The total mass of risk set at time, t, is

$$R^*(t) = \sum_{j=1}^{n}\left[\frac{a_j}{m_j^*}\sum_{i=1}^{m_j^*} I\,(T_{ji} \geq t)\right] \quad \text{..... Eq. 17}$$

The survival function can be estimated as

WC Estimator	$\hat{S}(t) = \prod_{j=1}^{n} \left\{ \prod_{i:T_{ji} \le t} \left[1 - \frac{d^*(T_{ji})}{R^*(T_{ji})} \right] \right\}$	..... Eq. 18

However when subjects are sampled from a target population, recurrent events are observed during the follow-up period and the initial occurrence of the event is not the requirement for recruitment, it is possible that some subjects may not experience any events during the follow-up period. Unfortunately, neither the PSH, nor the WC estimator is appropriate; in this case, both parametric and nonparametric approaches should be intended.

Although the above nonparametric approaches are widely used without assumption concerns, they provide no inference about group difference; they can only deal with categorical factors (continuous factors have to be categorized before analysis if they are involved); the approaches can only deal with a limited number of factors via stratification; they can estimate survival probability, but relatively low power for model prediction compares to parametric or semi-parametric models. Therefore, typical nonparametric approaches cannot be used for prognostic factor detection if there are more than a few categorical factors or if one or more continuous factors are involved. There are ways to get around some of these drawbacks by combining the nonparametric models with machine learning techniques, such as random survival forest[39, 40] (RSF), recursive partition[4] and support vector machine (SVT)[41]; however the results from the analysis are subject to random variation due to the nature of machine learning. Moreover, the nonparametric models are relatively less efficient for predictions. Thus it is still necessary to consider parametric or semi-parametric models for better inferences, mathematical simplicity and attractive properties.

2.3.2 Parametric Models of Survival Analysis

Parametric model is the one of the frameworks for implementation of multivariate regression analysis[42], in which the link function is expressed as a linear combination of all covariates (including the original factors or transformed factors, nonlinear form of the factors or transformed factors, interactions and etc.); the coefficients can be estimated via maximizing the joint likelihood of the link function. For parametric survival models, the

most commonly used distributions for survival time include exponential[43], Weibull[43], Gompertz[44], log-logistic and log-normal distributions. For most of these parametric models (except for exponential model), the survival time is changing over time following the distributions, thus these models are also called accelerated failure time (AFT) models.

2.3.2.1 Exponential Distribution and Exponential Survival Model

The exponential model[43] is one of the most popular parametric models based on exponential distribution; it is the simplest and possibly one of the most important model for survival analysis. Epstein et al. (1953) was among the first to introduce the exponential distribution for estimating the parameters for singly censored data; again in 1958, Epstein further discussed the justification of the exponential distribution assumption. Since then, the exponential model has continued to play an important role in survival analysis.

Exponential distribution is also one of the few distributions for which the estimator of its parameter has a closed-form solution when censoring is present. The estimator is a function of the number of events observed and the total exposure duration. This model is often used to analyze events which may have occurred "at random in time"[45].

In practice, the distribution is the basis for exponential proportional hazard model and is related to the extreme-value distribution; specifically, *T*, follows an exponential distribution with a constant hazard rate over the lifetimes of subjects. With the distribution assumption, the hazard, cumulative hazard, survival, density and cumulative distribution functions, all have simple forms; the mean and median survival time can be easily derived.

When $\lambda(t) = \lambda$ (fixed over time), then

$$\lambda(t|X) = \lambda(t)\exp(X\beta) = \lambda\exp(X\beta) \quad \text{.....Eq. 19}$$
$$\Lambda(t|X) = \Lambda(t)\exp(X\beta) = \lambda t\exp(X\beta)$$
$$S(t|X) = \exp[-\Lambda(t)]^{\exp(X\beta)} = \exp(-\lambda t)^{\exp(X\beta)}$$

where $\Lambda(t)$ is the cumulative hazard function; $S(t|X)$, is the probability of survival beyond time t given the values of the covariates, which can also be written as

$$S(t|X) = S(t)^{\exp(X\beta)} \quad \text{.....Eq. 20}$$

where the baseline survival probability $S(t) = \exp(-\lambda t)$.

The log-hazard, log-cumulative-hazard can be linearized with respect to $X\beta$ using the following identifies

$$\log \lambda(t|X) = \log \lambda(t) + X\beta \quad \text{..... Eq. 21}$$
$$\log \Lambda(t|X) = \log \Lambda(t) + X\beta$$

Thus, these can be analyzed using a regression analysis and the parameter λ is the antilog of the intercept. The expected failure time and median failure time is as follows

$$E(T|X) = 1/[\lambda(t)\exp(X\beta)] \quad \text{..... Eq. 22}$$
$$\mathrm{T}_{0.5}|X = \log 2/[\lambda(t)\exp(X\beta)]$$

This model has a property that the future lifetime of a subject remains the same no matter how long he stays in a study. This ageless property has made it a poor choice for modeling survival data from medical research except over a short period of time.

2.3.2.2 Weibull Distribution and Weibull Survival Model

Weibull model[43] is another parametric survival model based on Weibull distribution, which is a more generalized distribution. Unlike exponential model, this model does not assume a constant hazard rate; thus Weibull model has broader applications. The Weibull distribution was first introduced by Weibull in 1939 and its application to various failure time examples were discussed again by Weibull in 1955. Since then, the model has been broadly used for analyses of reliability and disease mortality.

The general form of the Weibull model is presented in Eq. 23; the distribution has two parameters, α and γ. The parameter γ determines the shape and the parameter α determines the scale of the distribution curve, therefore γ and α are called the shape and scale parameters, respectively. When $\gamma = 1$, the model becomes the exponential model and the hazard remains constant over times; when $\gamma > 1$, the hazard increases with time; and when $\gamma < 1$, the hazard decreases with time. Thus, Weibull model may be used to analyze survival data from a population with increasing, decreasing or constant risk.

The Weibull PH regression model is defined by the following:

$$\lambda(t|X) = \alpha\gamma t^{\gamma-1}\exp(X\beta) \quad \text{..... Eq. 23}$$
$$\Lambda(t|X) = \alpha t^{\gamma}\exp(X\beta)$$
$$S(t|X) = \exp(-\alpha t^{\gamma})^{\exp(X\beta)}$$

For numerical reasons, sometime it is advantageous to write the Weibull PH model as

$$S(t|X) = \exp[-\Lambda(t|X)] \quad \text{..... Eq. 24}$$

where $\Lambda(t|X) = \exp(\gamma \log t + X\beta)$.

The expected and median failure times are as follows

$$E(t|X) = 1/[\lambda(t)\exp(X\beta)] \quad \text{..... Eq. 25}$$

$$\mathrm{T}_{0.5}|X = \left\{\frac{\log 2}{[\alpha \exp(X\beta)]}\right\}^{1/\gamma}$$

The model can be diagnosed with graphical display, a linear relationship between $\log[-\log S(t)]$ vs. $\log t$ is a strong evidence for Weibull models. Another feature of the Weibull model is that it has both AFT form and proportional hazard form.

2.3.2.3 Gompertz Model

Gompertz distribution was introduced by Benjamin Gompertz in 1825; it was originally introduced for time series analysis. Since then, it has been a choice for modeling biological and survival data.

Gompertz survival model[44] is a 2-parameter survival distribution; the hazard function has the form, $\lambda(t|X) = \alpha \exp(\gamma t)$ where $\log \alpha = X\beta$. The log hazard is a linear function of the survival time with the form, $\log[\lambda(t)] = X\beta + \gamma t$. If the shape parameter, $\gamma > 0$, the hazard is monotonically increasing; if $\gamma = 0$, it is the exponential model; if $\gamma < 0$, the hazard declines monotonically.

This model is not commonly used, because the parameter estimation procedures, such as regression or maximum likelihood estimation (MLE) require knowledge of the actual lifespan for estimation of the parameters to be successful, which is not very realistic in analysis of survival data.

2.3.2.4 Gamma Distribution and Generalized Gamma Survival Model

In 1947, Brown et al. used a gamma distribution, a generalized form of both exponential and chi-square distribution, to describe the lifetime of glass tumblers served in a cafeteria; later in 1958, Birnbaum et al. used it as a statistical model to study the life length of various materials. Since then the model has been frequently used for reliability and survival problems. The model is intuitive for analysis of multiple-stage events or

failures; recently, it has been used as a distribution function of the random effect for clustered or multiple-event survival models.

In 2001, Pan W. discussed an AFT model with gamma frailty[46] to account for possible correlations and heterogeneity of multiple failure times. An EM-like algorithm was adapted for estimation. Several simulation studies were performed to compare the performance of this model with the other models assuming independence among multiple events.

This model was built to study multiple stage events or multiple failures per subject. The failure should take place in γ stages or as soon as γ sub-failures occurred. T_i is the time between the start of the ith stage, the ith sub-failure occurred and the $i+1$ sub-failure occurred. Then the total survival time, $T = \sum_i^{\gamma} T_i$, follows a gamma distribution, if the times $T_1, T_2, T_3, \dots, T_\gamma$ are independently distributed following an exponential distribution with the probability of $\lambda \exp(-\lambda t_i), i = 1, 2, \dots, \gamma$. Once the gamma distribution is obtained, the following can be derived easily following Eq. 2, Eq. 3 and Eq. 6, respectively.

$$f(t) = \frac{\lambda}{\Gamma(\gamma)} (\lambda t)^{\gamma-1} \exp(-\lambda t) \quad t, \gamma, \lambda > 0 \quad \text{..... Eq. 26}$$

$$F(t) = \frac{1}{\Gamma(\gamma)} \int_0^{\lambda t} u^{\gamma-1} \exp(-u)\, du$$

$$S(t) = \int_t^{\infty} \frac{\lambda}{\Gamma(\gamma)} (\lambda x)^{\gamma-1} \exp(-\lambda x)\, dx$$

2.3.2.5 Other AFT Model

Log-normal distribution[47, 48, 49] is another frequently used distribution for time-to-event data. In the simplest form, the lognormal distribution can be described using a normal distribution for the logarithm of the time, T; i.e., $\log T \sim \text{Normal}(\mu, \sigma)$. In 1879, the distribution was first introduced by McAlister et al.; in 1945, Gaddum et al. reviewed several applications for biological survival problems; in 1949, Boag et al. used this distribution in a cancer research to model survival times. Later, it was noted that the distribution of patient age at the onset of Alzheimer's disease and the distribution of survival time for several other diseases, such as Hodgkin's disease and chronic leukemia, could be approximated by a log-normal distribution.

Since then, this distribution has become more popular for survival analysis in part due to its theoretical relationship with normal distribution and in other part due to its practical approximation of survival time for certain disease[50, 51]. Similar to the normal distribution, log-normal distribution can be described using two parameters, μ and σ (see Eq. 28 for details). The distribution is the basis for log-normal accelerated failure time (AFT) model. Recently, this model has been extended to study multiple event survival data; in 1999, Klein et al.[52] utilized this model to perform a random effect survival analysis for censored data.

Other than log-normal AFT models, there are other AFT models include log-linear[53], log-logistic[54], Weibull extreme AFT model[55] and etc. These parametric AFT models have offered a variety of hazard function to model survival data; the effect of the linear combination of the covariates is to accelerate or decelerate the survival time of the event f interest. A common feature of the AFT models is the underlying assumption of the probability distribution for the logarithm of the survival time, $\log(T)$; the assumption may be a too restrictive to satisfy for modeling the distribution of the survival time. However, the estimates of coefficients from the AFT models are robust; i.e. the parameter estimates of the existing covariates are not affected by the omitted covariates[56], nor are they affected by the choice of the probability distribution for the logarithm of time. Moreover, it has been argued by Wei (1992)[57] and Cox (1997)[58] that the AFT model is more intuitively interpretable than PH model.

Recently, more complex models have been developed on the basis of the discussed AFT to account for random effects, correlated multiple or recurrent events, competing events and survival data with referral bias and etc.

In 2004, Lambert et al.[59] described a mixture of AFT models with a shared random effect to evaluate explanatory factors with adjustment of clusters effect of the investigator centers following kidney transplants. Different distributions for the random effects and the baseline hazard functions were considered; a flexible AFT model was proposed to account for both the short-term and long-term frailty effect for the hazard function. The model was then evaluated with the transplant survival data.

In 2013, Wang et al. described an AFT model to adjust for referral bias[60] when evaluating prognostic risk factors for progression in hepatitis C. In the paper, they

presented a study, of which the subject recruitment was significantly biased due to referral preference; subjects with more rapid disease progression were more likely to be recruited with liver clinics than any other clinics. A parametric random effect AFT model with adjustment of the correlation between the time referred to the clinics and the time to the development of cirrhosis was employed.

Another AFT model was described by Dang et al. (2013) for dealing with censored survival data with competing risks[61]; the model combined a mixture of AFT models for competing risks within a cluster weighted modeling framework; each competing risk was weighted by the cluster of the occurrence. They also used log-normal AFT model with alternating expectation conditional maximization algorithm for parameter estimation and bootstrap method for standard error estimation.

Below are the general forms of the AFT models; the log-logistic, log-normal, and Weibull extreme AFT models, can be derived by substituting the function ψ with the corresponding distribution functions, respectively.

General Form

$$S(t|X) = \psi\{[\log(t) - X\beta]/\sigma\} \quad \text{..... Eq. 27}$$
$$T_{0.5}|X = \exp(X\beta + \sigma\psi^{-1}(0.5))$$

where ψ is any standard survival distribution function.

Log-Normal Model

$$\hat{S}(t|X) = 1 - \phi\{[\log(t) - X\beta]/\sigma\} \quad \text{..... Eq. 28}$$
$$\hat{T}_{0.5}|X = \exp(X\hat{\beta})$$

Log-Logistic Model

$$\hat{S}(t|X) = \frac{1}{1 + \exp\{[\log(t) - X\beta]/\sigma]\}} \quad \text{..... Eq. 29}$$
$$\hat{T}_{0.5}|X = \exp(X\hat{\beta})$$

Similarly, the AFT equivalent of Weibull model can be obtained from the extreme value distribution (Eq. 23), by replacing γ with $1/\sigma$, and the median survival time can be obtained by replacing $[\alpha \exp(X\beta)]^{-1/\gamma}$ with $\exp(X\beta)$.

Weibull Extreme AFT

$$\hat{S}(t|X) = \exp\left[-\exp\left(\frac{[\log(t) - X\beta]}{\sigma}\right)\right] \quad \text{..... Eq. 30}$$
$$\hat{T}_{0.5}|X = [\log(2)]^{\sigma}\exp(X\hat{\beta})$$

Again, a graphic display is also the best diagnosis tool for parametric AFT models; for example, log-normal AFT model can be diagnosed by plotting the probability density of $\log T$; log-logistic AFT model can be diagnosed by detection of a linear relationship of

$\log\{S(t)/[1-S(t)]\}$ vs. $\log t$; Weibull extreme AFT model can be confirmed with detection of linear relationship between $\log[-\log S(t)]$ vs. $\log t$.

Comparing to Cox proportional hazard model, the AFT models are flexible, but they need additional verification for the distribution assumption for the baseline hazard, they are more complex, and needs numerical computations to obtain parameter estimates.

2.3.3 Semi-Parametric Survival Analysis

Cox proportional hazard (PH) model was first introduced by David Cox in 1972[62] to study age-specific failure rate, with a list of explanatory variables. Since then, the model has been adapted to a broad range of applications to health science, medical research, epidemiology and engineering industry. It is by far the most popular semi-parametric model[63, 64, 65, 66], for generalized survival regression analysis, where the regression coefficients are estimated with partial log likelihood. It allows for testing the difference between survival probabilities among groups of subjects while allowing for other prognostic factors; the group difference in survival probabilities is actually measured by hazard ratios. It makes a parametric assumption concerning the proportional effect of covariates with respect to the hazard ratio, but makes no assumption about the probability distribution of the baseline hazard. The outcome variable is the hazard rate as defined with Eq. 4 in section 2.2. The model assumes no probability distribution of the baseline hazard function but it does assume constant baseline hazard function, i.e., no time-dependency (such as time by covariate interactions).

Cox PH model is built on the rank ordering of the failure and censoring times, thus it is less affected by outliers in the event times than the parametric models. A second advantage of Cox PH model over other models is its ability to adjust for factors that are not modeled; hence, factors that are too difficult to model or do not satisfy the PH assumption, can be adjusted as the form of stratification factors. Thus, Cox PH model is efficient and robust; even when all assumptions for parametric models are satisfied, Cox PH model is still as efficient as parametric models.

When the proportionality assumption of Cox PH model is not met, extensions of Cox PH models are developed to relief the violation of the assumption. When recurrent or multiple events are of primary interest, various extensions of Cox PH models, such as Wei, Lin and Weissfeld (WLW)[67], Andersen-Gill (AG) [68], Prentice, Williams and

Peterson (PWP)[69] and etc., have been developed to handle clustered or multiple event data.

In the past decade, statistical learning approaches have been advancing rapidly; they were developed by applying machine learning techniques on top of the existing survival models or statistical testing procedures with the intension to overcome the limitations or restrictions of the typical statistical models. However, most of these were developed for regression and classification problems; and little has been done for survival data, and very few success stories have been reported for survival data. In this research, typical statistical approaches, multiple Cox regression models will the implemented and evaluated; the following statistical learning approaches will also be developed and assessed, including random survival forest[70, 71, 72] (RSF), generalized regression method[43] based on Cox model, regularized or penalized regression[73] method based on Cox model[74], derived input Cox model, such as principal component Cox regression[75] and partial least squares Cox regression models. These approaches will be evaluated based on two studies (one simulation study and one real world case study); the model performance will be assessed and compared.

2.3.3.1 Cox Proportional Hazard (Cox PH) Model

The Cox PH model can be commonly expressed as,

$$\lambda(t|X) = \lambda(t)\exp(X\beta) \quad \text{..... Eq. 31}$$

The Cox PH model can also be written in terms of the cumulative hazard and survival functions:

$$\Lambda(t|X) = \Lambda(t)\exp(X\beta) \quad \text{..... Eq. 32}$$
$$S(t|X) = S(t)^{\exp(X\beta)}$$

Thus the model can be linearized with respect to $X\beta$ using the following

$$\log\lambda(t|X) = \log\lambda(t) + X\beta \quad \text{..... Eq. 33}$$
$$\log\Lambda(t|X) = \log\Lambda(t) + X\beta$$

No assumptions were made for the baseline hazard function, $\lambda(t)$, or the baseline cumulative hazard function, $\Lambda(t)$; the baseline hazard is not needed for estimating the hazard ratio, formulated as $\lambda_1/\lambda_2 = exp[(X_1 - X_2)\beta]$, which only involves the relative

hazard function of $exp(X\beta)$; thus partial log likelihood function can be utilized to obtain the regression coefficients of β.

$$L(\beta) = \prod_{Y_i \text{ uncensored}} \frac{\exp(X_i\beta)}{\sum_{Y_j \geq t_i} \exp(X_j\beta)} \quad \text{..... Eq. 34}$$

$$\log L(\beta) = \sum_{i=1}^{k} \{X_i\beta - \log[\sum_{Y_j \geq t_i} \exp(X_j\beta)]\}$$

where k denotes the total number of unique failure times.

Since the model does not need to make assumptions for baseline hazard, and the partial log likelihood only involves the relative hazard function, such model does not have intercept parameter; the reference category omitted from the model will go into the underlying hazard function. Thus the hazard function for the reference category cannot be directly estimated.

The Cox PH model is convenient, flexible and often more powerful than other parametric or nonparametric models, however it has often been improperly used. When it is used for prognostic factor analysis, all prognostic factors are often included in their original linear form in the model without even considering interactions. Even if interactions are considered, only second order interactions with prior knowledge are included. Although there is no reason to believe higher order interactions do not exist. Additionally, forcing all factors to have a linear relationship with log hazard or log cumulative hazard may be convenient in formulating the model, but the linear relationships are rarely legitimate in practice[76]. Furthermore, the last but the most important step, assessments of proportionality assumptions are often skipped; instead, the Cox PH model is often applied without consideration of time-dependent covariates. For such model, only constant baseline hazard is allowed. In other words, the relationship of the covariates to the log hazard is assumed to be fixed at all values of t since $log[\lambda(t)]$ is separated from $X\beta$, which may not always be reasonable in clinical settings. Hence, inferences obtained from such models may be biased; predictions based on such models may have significant variance.

2.3.3.2 Cox PH Model with Time-Dependent Covariates

As discussed previously in section 2.3.3, typical Cox PH models does not adjust for time-dependent covariates, i.e., the values of all covariates are determined at the point when the follow-up begins on each subject (at time 0) and the values do not change over the course of the follow-up. However, in clinical settings, there may be situations in which the values of factors change over time, such as the level of AST/ALT may change over the course of Hepatitis. In these cases, the hazard function will depend more on the current values of the covariates than on the value at time 0. To handle similar situations, Abrahamowicz et al. (1996)[77] used regression splines to model the hazard ratio as a flexible function of time; Herndon et a. (1990) [78] used a restricted cubic spline function of time as the baseline hazard in a parametric proportional hazard model to account for time-dependent covariates ; Hess (1994) [79] also suggested to include a cubic spline function of time to adjust for non-proportionality and nonlinearity of time-by-covariate interaction.

To account for the time-varying covariates, extensions of the Cox PH model has been proposed allow for interactions between covariates and $\mathrm{g}(t)$, where g is a function of time. The extension of the Cox PH model is formulated as

$$\lambda(t|X) = \lambda(t)\exp\left(X_{(1)}\beta_{(1)} + \mathrm{g}(t)^T X_{(2)}\beta_{(2)}\right) \quad \text{..... Eq. 35}$$

where the covariate matrix X, is a combination of two parts, the 1st part, $X_{(1)}$ has all time-independent covariates, and $X_{(2)}$ has all time-varying covariates.

2.3.3.3 Cox-based Models for Recurrent Events

Besides time-to-single-event, researchers may also be interested in correlated event times due to multiple events or event recurrence. Multiple event survival data occurs when each subject may have one or more correlated events. Multiple events can be further classified into two categories, event recurrence and competing or multiple-type of events. Recurrent event data is defined as repeated occurrences of the same type of events in the same subject, such as repeated asthma attacks; multiple types of events occurs when each subject may experience events of completely different types, such as repeated occurrences of cancer cells at different sites, which may be referred to as competing events.

One popular extension of the Cox model recurrences of the same type of event is the Andersen-Gill extension[68]. The model assumes that events occurred within the same subject are independent, thus the number of events, $N(t)$, occurs over the interval, $[0, t)$, is considered as a counting process; then $dN(t)$ over a small interval $[t, t + dt)$ will take the general form as $\lambda(t|X) = \lambda(t)exp(X(t)\beta)$. To further account for heterogeneity among subjects, a random effect intensity model is proposed as

$$\lambda(t|\eta, X) = \eta\lambda_0(t)\exp[X(t)\beta] \quad \text{..... Eq. 36}$$

where η is the subject level random effect. However, the independent event assumption is still very much stringent, which may not be practical in clinical settings.

Other similar extensions include Prentice, Williams and Peterson (PWP) [69] model (proposed in 1981). The model is more powerful when there are a fairly large number of subjects available for analysis. Since the model is built as a generalized regression analysis with consideration of covariates and preceding failure time history; partial likelihood function is derived for estimation of the regression coefficients. Additionally, PWP model does not assume independent events within the same subject, thus it is more favorable.

Vaida et al. (2000) [80] proposed a generalized proportional hazard model with random effects for clustered or recurrent-event survival data. The generalized frailty model accounted for the random effects in the log hazard ratio, just as the random effects are modelled for generalized linear and non-linear mixed models. In the study, the distribution of the random effects was assumed to be multivariate normal, but they also claimed it should work with other distributions. The regression coefficients, variance components and baseline hazard function were estimated using EM algorithm.

Another nested gamma frailty model, proposed by Rondeau et al. in 2006[81], was another generalized random effect parametric model for adjustment of unobserved heterogeneity of dependent observations within nested clusters from the dataset. The nested frailty model used a hierarchical clustering with two nested levels of random effects to model repeated infections of subjects from different hospitals.

Rondeau et al. (2008)[82] used a Gaussian frailty Cox model with additive random effects in a meta-analysis to combine the survival analysis results from different clinical trials. The model accounted for possible heterogeneity among different clinical trials; a

general correlation structure was considered for the random trials effects or the random treatment-by-trial interaction effects. The regression coefficients and hazard function were estimated using a semi-parametric "penalized marginal likelihood method" with a pre-specified variance-covariance structure for the random effects.

2.3.3.4 Cox Models for Competing Events

When recurrent events (of the same type) from the same subject are terminated by a major failure event (different type) and the terminal failure event may be correlated with recurrent events, then they are considered as competing events; the recurrent events are competing with the terminal event. There are a variety of ways to model the random effects for analysis of competing events. A popular choice was to use a marginal model for the recurrent events and a Cox model with time-dependent covariates summarizing the history of the recurrent events for the terminal event. This approach is referred to as the "selection model" defined by Little et al. [83] in 1995.

Alternatively, Little et al. (1995) also suggested a shared or correlated random effects to account for the association between the recurrent events and terminal event. When the primary interest is to characterize the recurrent events, the terminal event may be adjusted with a "pattern-mixture" model.

In 1997, Li and Lagakos [84] presented another example with marginal WLW [67] extension of the Cox model; in the study, the terminating event was considered as a censoring event for the recurrent event, and each occurrence time of the recurrent events was considered as the first occurrence for the next event (including both recurrent events or the terminal event, whichever was the first). In 2003, Ghosh and Lin [85] proposed a joint marginal Cox model to evaluate recurrent events with a correlated terminal censoring event.

Other methods based on counting process were also proposed. In 1998, Lancaster and Intrator [86] presented a joint parametric Poisson model to evaluate hospitalization experience with impatient repeated episodes of infections from HIV positive patients. The model used a subject-level frailty term to evaluate the recurrent failure event with adjustment of inter-subject and intra-subject correlations. Sinha and Maiti (2004)[87] considered a more general model, a hierarchical Bayesian framework, to adjust for recurrent events following a counting process and a dependent terminal events.

In 2002, Huang and Wolfe [88] suggested to utilize the informative censoring as terminal event for recurrent event survival data. In 2004, Liu et al. [89] presented a joint semi-parametric model with a shared gamma frailty effect to address the intensity functions of both recurrent events and terminal event. In these models, the random effects on recurrent events and terminal event were slightly different and the regression coefficients were estimated using Monte Carlo expectation maximization (EM) algorithm.

Rondeau et al. (2007) [90] presented an analysis of repeated occurrences of follicular lymphomas for subjects who could be terminated by a terminal event. In the study, the terminal event was correlated with the recurrent events from the same subject; the nominal assumption of non-informative censoring of the recurrent events by a terminal event was violated. A joint Cox frailty model was proposed to obtain the unbiased parameter estimates for both the recurrent events and the terminal event.

Recurrent Event	$r(t\|X,v) = v\mathrm{r}_0(t)\exp(X\beta)$	..... Eq. 37
Death	$\lambda(t\|X,v) = v^{\alpha}\lambda_0(t)\exp(X\gamma)$	

where $r_0(t)$ and $\lambda_0(t)$ were the recurrent and terminal event baseline hazard function, β and γ were the regression coefficient vectors for X; the random effect vector, $v \sim \Gamma(\frac{1}{\theta}, \frac{1}{\theta})$.

2.3.3.5 Random Effect Cox Models (Shared Frailty)

Mauguen et al. (2013)[91] proposed a shared frailty Cox model to account for gamma distributed or log-normal distributed random effects of clusters; clustered survival times were evaluated. Cox model was used for estimation the random effect Cox model; marginal log likelihood function was employed. The regression coefficients were estimated through the maximization of the penalized marginal log-likelihood or through the maximization of the log-likelihood using the robust Marquardt algorithm[92].

2.4 Generalized Regression Analysis for Survival Data

$Y(t|X)$ is a generic notation for the link function, which can be a matrix of time-to-single-event or a matrix of time to competing multiple types of events. For Cox PH model, it is the log HR; note that for Cox PH model, no assumption is made for baseline hazard, $\lambda(t)$, and log HR was the link function for the generalized regression analysis.

A generalized regression analysis has the following forms.

General Form	$Y(t\|X) = X\beta$	..... Eq. 38
Vector Form	$y(t\|x) = \beta^T x$	

The regression coefficient, β, can be estimated with linear algebra or can be estimated through maximization of the partial log likelihood.

Regression coefficient	$\hat{\beta} = (X^T X)^{-1} X^T Y$	..... Eq. 39
Predicted Value of Y	$\hat{Y} = X\hat{\beta}$	

The above solution needs the inverse of the matrix $X^T X$, which however could be singular (non-invertible), as such the generalized regression may not have an estimable solution. There is a way to get around with the help of partial least squares[93, 94] regression approach, which was first introduced as a machine learning technique.

A generalized random effect model should include terms for typical fixed effect and terms for random effect from a cluster or group, therefore the transformed outcome for subject j who is from cluster l will have to form:

$$Y_{lj} = \beta^T x_j + U_l + \epsilon_{lj} \quad \text{..... Eq. 40}$$

where x_j is a vector with values of all covariates for subject j; U_l is the cluster-specific random effect, which measures the difference between the average of cluster l and the average of entire population; ϵ_{lj} is the subject level random error.

2.5 A Review of Machine Learning Techniques

Principal components analysis (PCA) was first invented by Pearson Karl (1901)[95] as an analogue of the "principal axes theorem" in mechanics; it was later independently developed (and named) by Harold Hotelling in the 1930s[96]. The approach is a procedure to transform possibly correlated continuous covariates into orthogonal space, such that components may become linearly uncorrelated. The uncorrelated components can be used as covariates for regression analysis and thereafter to select best subset of covariates. A brief description of the approach is presented in section 3.3.3.1.

However, the approach has many disadvantages. Components are constructed independent of the response variable, thus there is no guarantee that the constructed

components are correlated with the response and because the relationship between the original factors and the response are linked by the constructed components, it is difficult to interpret the results without transformation back to the original factors. Moreover, the constructed components still rely on typical statistical models to estimate, thus the principal component regression may still be non-estimable due to too many components.

Correspondence analysis (CA) proposed by Hirschfeld (1935)[97] was a similar multivariate statistical model to adjust for correlated categorical variables. It creates orthogonal components based on the contingency table from the categorical variables. Another extension, multiple correspondence analysis (MCA) is an analysis technique for nominal categorical data, to detect and represent the underlying structures in a dataset. For simplicity, both will be called correspondence analysis in the paper.

Similar to principal component Cox regression, Eric et al. (2004) proposed a semi-supervised principal Component analysis; the idea was to build only a few components to cluster small number factors that were highly correlated with response. The components that were highly correlated to the response should be much more important than those that were not correlated to the response; the components were constructed with several important factors with high predictive powers. This approach has overcome several of the major disadvantages of a principal component analysis; if none of the factors are really correlated with response, this model is no worse than a principal component analysis; but it can handle the case when the number of components is more than the total number of observations; if there are a few factors highly correlated with the response, the model can correctly cluster them into smaller components using a likelihood ratio test (unsupervised part); unlike the principal component analysis, the model is not completely based on parametric approach, therefore it guarantee to obtain a model fit.

Ordinary least square (OLS) analysis often does poorly in predictions; the idea of penalization has been proposed to improve the prediction of OLS, such as ridge, lasso and least angle regression. The ridge regression was first introduced into statistics by Marquardt et al (1970) [98] borrowing the regularization idea from Andrey Tikhonov (1943)[99] and Philips et al. (1962) [100]; the model can be clearly described using a loss function of the residual sum of squares with a $L2$ penalization ($L2$-norm of the coefficients, $\lambda\beta^T\beta$). As a continuous shrinkage method, ridge regression achieves better

prediction through bias–variance trade-off. Unfortunately, this approach usually keeps most of the covariates in the model if not all, and the variable selection can produce a "sparse model", which is extremely unstable because of its "inherent discreteness" (Breiman et al. 1996[101]). In 2012, Jelle J. Goeman[16] introduced the ridge regression into survival analysis.

$$\mathrm{L}^{\mathrm{ridge}} = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \quad \text{..... Eq. 41}$$

$$\hat{\beta}^{\mathrm{ridge}} = (X^TX + \lambda I)^{-1}X^Ty$$

Or $$\hat{\beta}^{\mathrm{ridge}} = \mathrm{argmin}\left\{\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda\sum_{l=1}^{v}\beta_l^2\right\}$$

Lasso regression is another penalized regression model proposed by Tibshirani (1996)[102]; in 2010, Jelle Goeman[17] introduced this approach to survival analysis, and later in 2012[16], he compared the performance of the lasso with the ridge Cox regression models. Similar to ridge regression, the loss function is defined as the residual sum of squares with an $L1$-penalty on the regression coefficients, $\lambda\sqrt{\beta^T\beta}$. One of the prime differences between Lasso and ridge regression is the penalization effect; for ridge regression, as the penalty increases, all parameters are shrinked toward zero but never reaches zero; for lasso regression, the increase in the penalty term will drive more parameters to zero and only the covariates with nonzero coefficients left in the model are selected. Owing to the nature of lasso regression, it does both shrinkage and variable selections simultaneously.

$$\mathrm{L}^{\mathrm{lasso}} = \frac{1}{2}\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda\sum_{l=1}^{v}\|\beta_l\| \quad \text{..... Eq. 42}$$

$$\hat{\beta}^{\mathrm{lasso}} = \mathrm{argmin}\left\{\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda\sum_{l=1}^{v}\|\beta_l\|\right\}$$

where $\hat{y}_j = \sum_{l=1}^{v} x_{jl}\beta_l$.

For model performance, Tibshirani (1996) and Fu (1998)[103] compared the penalized models including the lasso, ridge and bridge regression; they found that none of these models uniformly dominated the other two in terms of predictions accuracy. However, as variable selection becomes increasingly important in modern data analysis, lasso has become much more appealing.

Although lasso regression has shown some success, but it is limited in several situations; when there are more covariates than the number of observations, lasso will be

limited in variable selection; for highly correlated covariates, lasso regression tends to randomly select one from the group of correlated variables. When there are high correlations among covariates, the performance of lasso regression can be dominated by ridge regression [104].

To resolve the drawbacks that both ridge regression and lasso regression encounter, another shrinkage approach, Elastic-net regression was proposed by Zou et al. (2005)[104] for continuous outcomes. The penalization term includes both the $L2$ norm of the regression coefficient from ridge regression and the $L1$ penalty from Lasso regression (see section 3.3.2.1 for details). In terms of performance, elastic-net regression often outperforms both lasso and ridge. In terms of variable selections, the model encourages grouping effect for correlated factors; strongly correlated covariates tend to be all-in or all-out of the model together. The approach is especially helpful when the number of covariates is much more than the total number of observations. In 2004, Efron et al. [105] proposed an algorithm based on the "least angle regression" (LARS) to automatically estimate the penalization terms efficiently, which made this approach even more attractive.

Partial least squares (PLS) regression was first introduced by a Swedish statistician Herman Wold, and it was fully developed by his son, Svante Wold (2001)[106]. The idea was to project all factors to the hyper-planes in the direction of the response. Today, PLS regression model is most widely used in chemometrics, bioinformatics and many other areas. In 2002, an extension of the PLS approach, orthogonal projections to latent structures (OPLS), was proposed by Trygg et al.[107]; the approach separated continuous covariates into predictive and uncorrelated components, which may be used for prognostic factor detection. This new extension improves the diagnostics, interpretability and visualization, but it does not affect the model predictions. Details of the algorithms will be provided in section 3.3.3.2.

Another type of approach for regression and classification problems is the decision tree learning [108], which utilizes a decision tree as a predictive model to "map observations to their target value". In the tree structure, leaves represent classes or subgroups, branches represent conjunctions of features that lead to the leaves (classes). Recursive partition[109] was proposed by Leo Breiman in 1984; this approach recursively

split the data into subsets based on the response, and it is completed when further splitting does not add any value to the predictions of the outcome (detailed algorithm will be described in section 3.3.1.1). However, the decision tree based learning approaches are well known for low bias and high variance; they tend to overfit the data (large variance).

In 1994, Leo Breiman proposed a bootstrap aggregating (Bagging)[110] to improve the stability and accuracy of decision tree approach; in this approach, new training set is repeatedly sampled with replacement from the original training set, decision is voted within each bootstrapped training set, prediction is made by averaging all decisions, which can help to achieve better stability and accuracy. In 2001, Leo Breiman[111] proposed another Bagging based approach, Random forest. The approach has achieved substantial improvement over the traditional decision-tree based learning; in this approach, a large collection of de-correlated trees were built, all trees are averaged for the final decisions or results. So far, the approach has been widely used for assessing continuous and categorical outcomes. (Detailed algorithm will be discussed in section 3.3.1.2). However, it is difficult to interpret the result from the analysis of random forest because of the "black-box" prediction.

To combine the advantage of accuracy with reasonable interpretability, recently, Song et al.[112] proposed a random generalized linear model, which shares the same advantages of random forest model with a forward-selected generalized linear model, it is claimed to have excellent predictive accuracy, feature importance measures and outstanding interpretability.

Chapter 3. Research Methodology

To improve the efficiency of typical Cox PH models, a systematical process was proposed to automatically assess the model assumptions, suggest reasonable solutions to resolve violations if detected and/or relax the model assumptions when necessary. The typical Cox PH models were implemented with three options, generalized Cox linear regression model /or (in short) Cox linear model, multivariate Cox regression models[43] with restrictive cubic spline (*RCS*) transformations /or (in short) Cox model with *RCS* transformations) and/or with fractional polynomial transformations (*FP*) /or (in short) Cox model with *FP* transformations.

For generalized Cox linear regression model, all factors were included in their original linear forms, see section 3.1 for details; for multivariate Cox regression model with *RCS* and *FP* transformations, all continuous factors were included using *RCS* and *FP* transformation respectively, the interaction terms for either model were constructed using the corresponding transformations (see section 3.2 for details). The two transformation options for multivariate Cox regression are useful for modeling the nonlinear covariates. Unfortunately, they also add extra burdens to the analysis, especially for survival data, for which only the failure events of interest are contributing to the analysis. With the extra terms from the *RCS* or *FP* transformation, the Cox regression model have to spent more degree of freedoms for the analysis; thus it becomes quite cumbersome when there are more factors to be analyzed; furthermore, when the number of factors is close to or more than the number of events available, the typical Cox regression models become non-estimable.

As mentioned in section 2.3.3, Cox PH model is very flexible and robust, even for survival data that satisfies complete parametric models; however it does have somewhat stringent assumptions. Nonparametric approaches on the other hand can only deal with a limited number of categorical factors with relative low prediction power. To overcome the disadvantages of the typical nonparametric and semi-parametric approaches, nonparametric random survival forest[5, 39, 113] approaches (see section 3.3.1.2 for details) were introduced, which should be able to handle many more predictors including both continuous and categorical factors, and it should work better than the typical Cox PH models when the number of factors is no less than the total number of events or the

survival time does not follow any known survival distributions; other than the nonparametric approaches, a semi-parametric approach, elastic-net was developed based on Cox model with the intension to achieve variable selection and obtain inferences simultaneously, with reasonably good prediction performance in cases; this approach should be able to handle cases when there are too many predictors to be estimated with typical Cox PH model; aside from the elastic-met Cox regression, an partial least squares Cox PH model was developed to adjust for uncollected covariates for highly correlated survival data.

Then the other approaches based on machine learning techniques should become useful, including random survival forest, penalized Cox regression[104] (section 3.3.2.1) including lasso, ridge and elastic-net Cox regression models, derived input Cox regression models such as principal component Cox regression model (section 3.3.3.1) and partial least squares Cox regression model (section 3.3.3.2).

Additionally, the decision tree based learning, recursive partition[114] was also implemented in this research; it was only used to impute missing value for categorical variables in the survival data during the data preparation step (Section 3.3.1.1); it was not used for survival analysis due to its instability and unreliability.

Besides the above approaches, here are a few other tools implemented in this research. The least squares multiple regression incorporating optimum transformations were implemented for variable transformations[115, 116, 117, 118] and missing data imputations[119, 120] for continuous variables. Multicollinearity were assessed using Spearman's ρ^2 rank correlation[121, 122]; factors were clustered using hierarchical cluster analysis[123] via Hoeffding's D statistics[124]. Factors identified from hierarchical cluster analysis were processed using principal component analysis[125, 126, 127] (PCA) with maximum total variance (MTV)[128] method; variable reduction or cluster should be confirmed by the PCA. The clustered variables were used to construct principal components, which were then used to replace the original factors in the survival analysis; multicollinearity should be fixed consequently. The last, but the most important part of the research was to propose model performance statistics for assessment of prediction accuracy and prediction powers for different survival models; for this purpose, the

concept of prediction errors and time-dependent AUCs were defined (see section 3.11 for details).

3.1 Generalized Cox Linear Regression Model

For this model, all factors or composite factors as derived from section 3.7 were included as the covariates in the Cox model in their 1st order linear form, potential 2nd order interaction terms between factors (or composite factors) as confirmed from Section 3.8.4 were also included in the model; if composite factors were derived, they were used to replace the original factors in the model. The best model was then selected following the model selection procedures as described in Section 3.9; this model will also be referred to as Cox linear model (in short). The application was implemented with R/rms[129] package.

3.2 Multivariate Cox Regression Models with Nonlinear Transformations Including Restrictive Cubit Spline (*RCS*) and Fractional Polynomial (*FP*)

Multivariate Cox regression models with nonlinear transformations were implemented; the nonlinearity of the original factors (see section 3.8.3 for details on nonlinear transformations) or composite factors (see section 3.7 for details on variable clustering) was adjusted using two options, the *RCS* or *FP* transformations; all potential interactions were constructed between the transformed factors included in the model. Note that only continuous factors were transformed using restricted cubic spline (*RCS*) or fractional polynomial (*FP*) transformations (see section 3.8.3 for details). If interactions were confirmed following the instruction from section 3.8.4, the interaction terms should be constructed between all *RCS* forms (or *FP* forms) of the factors or composite factors involved in the interactions. For time varying covariates, time-dependent extensions of Cox PH model were intended (see section 3.8.5 for details). The multivariate Cox regression models with the two nonlinear transformation options were referred to as multivariate Cox regression models with *RCS* and *FP* transformations, or (in short) as the Cox model with RCS or FP transformations, respectively; time-dependent covariates or extensions may be considered if necessary.

Given survival data $(x_j^T, t_j), j = 1, 2, \dots, n$, where n is the total number of subjects, x_j^T is a row vector with all factors from subject j, $x_j^T = (x_{j1}, x_{j2}, \dots, x_{jp})$; $y(t_j|x_j)$ is the link

function corresponding to (x_j^T, t_j) from subject j. Then the analysis model is:

$$y(t_j|x_j) = \sum_{l \in G_{\text{lin}}} \beta_l x_{jl} + \sum_{l \in G_{\text{nonlin}}} f_l(x_{jl}) + \epsilon_j \quad \text{..... Eq. 43}$$

with $\epsilon_j \sim N(0, \sigma^2)$. $G = \{1, 2, \ldots p\}$, where p is the total number of factors. $G_{\text{lin}}, G_{\text{nonlin}} \subset G$; in particular, model selection is to identify the set $G_{\text{noe}} = \overline{G_{\text{lin}} \cup G_{\text{nonlin}}}$ of the covariates that are not included in the model (see Section 3.9 for details of model selection). The multivariate Cox regression model with *RCS* transformations were implemented with R/rms[129] package; the Cox model with *FP* transformation were implemented with R/mfp[130], coxph[131, 132] and rms[129] packages.

3.3 Machine Learning Techniques

As discussed in the previous chapter, the parametric or semi-parametric model should always have more power for making predictions and inferences; however in case the primary interest was to detect prognostic factors and predict the probability of future occurrence of failure event, without concerning inferences about the coefficients, classification approaches with consideration of centering information should be useful.

3.3.1 Tree Based Approach

There are many tree based approaches available, but in general they can be classified into two types; one type is regression-tree based approach, such as recursive partition and regression trees; these approaches are based on a single run of the data, therefore they have small bias but large variance and the variable selection process is not stable. Another type is bootstrap aggregation based approach, such as random forest; there are some random variations due to the nature of bootstrap, therefore they usually have relatively large bias but small variance compared to the regression-tree based approach. In this research, recursive partition was used to impute missing categorical variables and random survival forest was implemented for modelling the survival data.

3.3.1.1 Recursive Partition

The recursive partition approach was used to impute missing values of the categorical variables in the studies. Let f be some impurity function, then the impurity (or diversity) of node A is, $I(A) = \sum_{i=1}^{C} f(p_{iA})$, where p_{iA} is the proportion of subjects in node A with

class i and C is the total number of different classes. Node A is split based on the maximum reduction in impurity, $\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R)$, where A_L and A_R are the two offspring nodes from A. The recursive partition was implemented with R/rms[129] package.

3.3.1.2 Random Survival Forest (RSF)

Random forest[5, 39] learning is recently developed, which is believed to have much better stability and accuracy over most decision tree based learning. Random forest learning is implemented based on nonparametric bootstrap aggregation (Bagging); it was originally developed for regression analysis on continuous outcomes, such as linear regression, for logistic regression analysis on binary outcomes, or for multiple regression and multinomial regression on nominal or ordinal outcomes. Only recently, the approach has been applied to model survival outcome; though there are a few publications on this topic, still there are more to improve, especially for model prediction and prediction performance. This model is implemented through bootstrap, which does not rely on any distribution assumptions; additionally, the approach makes no assumption about the covariates, therefore there should be no need to worry about variable transformation, nonlinearity, multicollinearity or time dependency.

In this paper, two different version of random survival forest (RSF) were implemented. The first one was the random survival forest based on the log-rank test; one take-over advantage of this approach is that it could systematically detect interactions; this approach will be referred to as log-rank based RSF (LR-RSF) model onward. The second one was conditional inference based random survival forest[133, 134]; in this approach, forest trees are split based on the conditional probability, where multiple log-rank tests are computed at each start of the algorithm based on permutation test, thus it is expected to have better performance for highly correlated survival data; this approach is referred to as conditional inference based RSF (CINF-RSF) model onward. The test statistics can be obtained with the average of all possible values of the test statistic under rearrangements of the bootstrapped samples, thus it should have reduced overfitting and variable selection bias. For both approaches, the implementation algorithm was the similar.

Given a training set with N observations:

1. From $b = 1$ to B:
    a. Draw a bootstrap sample (sample with replacement) of size N;
    b. Grow a random forest tree for the sample, by recursively repeating the following steps for each terminal node of the forest tree, until the minimum node size n_{min} is reached.
        i. Select m variables at random from the p covariates.
        ii. Pick the best variable and split-point using the log-rank test.
        iii. Split the node into two offspring nodes.
2. Output the random forest trees $\{\hat{f}_b\}$ from the B bootstrapped sample.

To make a prediction at a new time point, t, based on B forest trees:
Survival probability: the predicted survival probability, $\hat{S}_{rf}^{B}(t) = 1/B \sum_{b=1}^{B} \hat{S}_b(t)$;
Survival status: Let $\hat{C}_b(t)$ be the predicted survival status from bth tree. Then the predicted survival status of all B forest trees: $\hat{C}_{rf}^{B}$ =majority vote $\left\{\hat{C}_b(t)\right\}_{b=1}^{B}$.

Log-rank based RSF model was implemented with R/randomSurvivalSRC[135, 136, 137] package and conditional inference based RSF model was implemented with R/party[138, 139, 140] package.

3.3.2 Shrinkage or Penalized Regression Analysis

There are several shrinkage or penalized regression models, such as ridge, lasso, least angle and elastic net regression models. In this paper, only lasso, ridge and elastic-net Cox regression were implemented. The lasso and ridge models were briefly described in section 2.5, and they are considered as special cases of elastic-net Cox regression model (Lasso can be considered as elastic-net with λ_2=0 from Eq. 44 or α=1 from Eq. 45 and ridge can be considered as elastic-net with λ_1=0 from Eq. 44 or α=0 from Eq. 45). Thus, in this section, only elastic-net regression model will be discussed.

3.3.2.1 Elastic-Net Regression

Elastic-Net regression model[141] is a regularized least-squares regression approach for variable selections; it bears some resemblance to ridge and lasso regression, with the only difference in the penalization term, which involves the $L2$ norm from ridge regression (see section Eq. 41 from Section 2.5 for details) and the $L1$ norm (see Eq. 42

from Section 2.5 or details) from lasso regression, where $L1$ and $L2$ are also known as the penalty terms or regularization parameters.

$$\mathrm{L}^{\mathrm{EN}} = \frac{1}{2}\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda_1 \sum_{l=1}^{v}\|\beta_l\| + \lambda_2 \sum_{l=1}^{v}\beta_l^2 \quad \text{..... Eq. 44}$$

$$\hat{\beta}^{\mathrm{EN}} = \operatorname{argmin}\left\{\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda_1 \sum_{l=1}^{v}\|\beta_l\| + \lambda_2 \sum_{l=1}^{v}\beta_l^2\right\}$$

where $\hat{y}_j = \sum_{l=1}^{v} x_{jl}\beta_l$ or $\hat{\beta}^{\mathrm{EN}}$ can be re-write as

$$\hat{\beta}^{\mathrm{EN}} = \operatorname{argmin}\left\{\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda\alpha \sum_{l=1}^{v}|\beta_l| + \frac{1}{2}\lambda(1-\alpha)\sum_{l=1}^{v}\beta_l^2\right\} \quad \text{..... Eq. 45}$$

where $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)/2$.

Lasso and ridge regression are just special case of this formula; if α=1, it becomes lasso regression; if α=0, it becomes ridge regression. The regression coefficients can be obtained by minimizing the loss function (Eq. 45), then the hazard, log hazard and log cumulative hazard can be obtained from Eq. 19, Eq. 20 and Eq. 21, respectively.

Unlike lasso regression which only chooses a few nonzero coefficients and ridge regression which tends to keep all covariates by shrinking all coefficients towards 0, which never reaches 0. The elastic-net model combines the strength of both ridge and lasso regression; the penalization terms, λ and α, are chosen to balance between the two with a cross validation (CV) step; for a fixed λ, as alpha changes from 0 to 1, the behavior of the model moves from ridge-like to lasso-like regression, increasing the magnitude of all non-zero coefficients. With α =0.95 or above, the elastic-net behaves similar to lasso regression. Variables are selected through the penalization terms; covariates are excluded if the regression coefficients are zeroes. All three penalized regression approaches were implemented with R/glmnet[142] package.

3.3.3 Derived Input Regression

3.3.3.1 Principal Component Regression

Principal component analysis is an unsupervised approach, since the construction of the components from the covariates is not based on the response. Once covariates are clustered into components based on principal component analysis (PCA); the selected

components can be used as covariates for Cox regression analysis; this approaches will be referred to as principal component Cox regression (or PCR) model in this paper. There are several versions of principal component analyses, however it was noted that different PCR models should yield similar results; in this research, only the covariance based PCR model was evaluated. The PCR model was selected via cross validation; coefficients of the principle components were estimated based on the selected PCR model (the coefficients of the original factors could be obtained by inverse-transforming the coefficients of the principal components), prediction of unseen data was made with the selected PCR model and prediction performance was assessed and compared with other survival models. For PCR analysis, the following steps should be followed:

Derive orthogonal components as $z_m = Xw_m$, where $m = 1, 2, \dots, M$, where $M \leq v$, v is the total number of covariates, M is the total number of components. Then

$$\hat{y}_{(M)}^{\text{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m \quad \text{..... Eq. 46}$$

where $\widehat{\theta}_m = \frac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle}$. This is the initial PCR model, component scores for further variable clustering can be obtained from this model; for assessment of the model fit, the performance of this initial model is compared with a tentative fit from a Cox PH linear model (see section 3.1 for details) and tentative fits from multivariate Cox regression models (see section 3.2 for details).

This PCR model can be processed to select the best subset of components ($\widehat{M}$). Based on the selected PCR model, the log hazard can be regressed on the selected components as the only covariates using ordinary least squares regression to obtain the regression coefficients (with dimension equal to the number of the selected components).

At last, the regression coefficient of the components can be inverse transformed to obtain the coefficients of the original covariates, using the selected PCA (or CA) loading matrix corresponding to the selected principal components.

$$\hat{\beta}^{\text{pcr}}(\widehat{M}) = \sum_{m=1}^{\widehat{M}} \hat{\theta}_m w_m \quad \text{..... Eq. 47}$$

3.3.3.2 Partial Least Squares Regression

Partial least squares (PLS)[106, 107] regression bears some resemblance to the principal components regression; both utilized orthogonal latent components to link the original factors with the response. However, unlike the principal component regression which constructs the components independent of the response, the PLS model tries to find the multidimensional direction of the explanatory covariates that explains the maximum variance in the direction of the response. The principal components can achieve dimensional reduction by projecting the covariates (though linear combinations) into uncorrelated components that bear minimum variance; by looking at the contribution of each component to the total variance of the covariates, the ones with the minimum contributions are considered as redundant, thus it achieves variable reduction without looking at the responses. Well, partial least squares regression shares the same thoughts of projecting the covariates into orthogonal components that can explain the maximum variance of the response; through the intermediate latent components, the fundamental relationships between the original covariates and the response can be elaborated. The theory of PLS was developed by Leo Brieman in 2001[5], since then this approach has earned it reputation in many disciplines. In this research paper, this approach was developed and implemented to model the survival data, since there had been many successful stories for predicting continuous and categorical outcomes with similar models.

Given a training set with N:

1. Standardize each x_l to have mean zero and variance of 1. Set $\hat{y}^{(0)} = \bar{y}\mathbf{1}$, and $x_l^{(0)} = x_l, l = 1, 2, \dots, \nu$.
2. From $m = 1, 2, \dots, \nu$
   a. $z_m = \sum_{l=1}^{\nu} \hat{\varphi}_{ml} x_l^{(m-1)}$, where $\hat{\varphi}_{ml} = \langle x_l^{(m-1)}, y \rangle$.
   b. $\hat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$.
   c. $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$.
   d. Orthogonalize each $x_l^{(m-1)}$ with respect to z_m:

$$x_l^{(m)} = x_l^{(m-1)} - \left[\frac{\langle z_m, x_l^{(m-1)} \rangle}{\langle z_m, z_m \rangle} \right] z_m, \; l = 1, 2, \dots, \nu.$$

3. Output the sequence of fitted vectors $\{\hat{y}^m\}_1^{\nu}$. Since $\{z_h\}_1^m$ are linear in the original x_l, so is $\hat{y}^m = X\beta^{\mathrm{PLS}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformation.

where $j = 1, 2, \ldots N$, N is the total number of subjects in the training set; $l = 1, 2, \ldots, \nu$, ν is the total number of covariates (including linear, nonlinear and interaction terms); $\bar{y}\mathbf{1}$ is used to denote a vector of constant, which is the mean of vector y.

3.4 Sample Size, Data Simulation or Data Selection

For survival analysis, not all information from survival data contributes to the analysis and parameter estimation; only information about event occurrence should provide useful information for analysis rather than censoring information. Thus, only the total number of events was of concern for parameter estimation. As a rule of thumb, without considering any interactions and non-linear effect, the total number of events had to be at least 10 times[43] more than the total number of factors to be included in the survival model. If pair-wise interactions or non-linear effects (such as polynomial terms or restricted cubic spline transformed terms) were suspected, much more events should be needed to reasonably estimate all coefficients from the survival models. Hereafter, covariate or covariate term will be referred to as the individual term included in the analysis model; it may be a term from the polynomial transformations (i.e., $x_1^3, x_1^2 \; and \; x_1$ are 3 terms for 3 degree polynomial transformation of x_1 or an interaction term (such as $x_1 : x_2$). As an example, if p factors [A, B, C …] are believed to be linear without any interactions, then there will be p terms in the model, then at least $10 \times p$ events are required for reasonable estimation of all p terms; for the same number of factors, if piecewise second order interactions exist among the 3 factors [A × B, A × C, B × C…] and factor A was transformed using a κ-knot restricted cubic spline (*RCS*) transformation A, [denoted as *RCS*(A, κ)], and all the rest of the factors are included in the linear form, then a total of $\nu = p + \binom{p}{2} + (\kappa + 1)$ terms should be included in the model, and 10 times more events are required in order to estimate all parameters from the model; if more than one factor needs *RCS* transformation, more covariates terms should be included, consequently more events are required. In the above example, inclusion of the

restricted cubic spline transformation for factor A in the survival model is only considered if one or more inflection points are detected of the residual plot: the minimum number of knots, κ, is determined by the number of inflection points (see section 3.8.3 for details).

Despite the size of the data, it is always easy to construct more terms than collection of more observations for analysis. To ensure enough data for analysis, in the simulation study, 2000 observations with 7 factors, including age, sex, race, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI) and treatment were simulated (see section 4.1 for detailed description of the simulation). For the real world case study, a 70-gene-signature breast cancer[143] data was downloaded from Netherlands Cancer Institute to evaluate metastasis-free survival; in which, the number of factors was than the total number of events available. In the circumstances, typical Cox PH models did not work due to singular design matrix, then all other available models were evaluated, including random survival forest (RSF), penalized Cox-regression models , principal component Cox regression and partial least squares Cox regression models.

3.5 Data Transformation and Normalization

For simulation study, a majority of the continuous factors were transformed using restricted cubic spline function or fractional polynomial function if needed; other common transformation technique were also intended, such as normalization, log transformation, exponential and polynomial transformation, etc. For factors that could not be transformed to satisfy model assumptions, they should be included as stratification factors; but since the intension was to build predictive survival models for the entire study, not to evaluate the models for each stratification, therefore this step was not considered in this research. In case it was necessary, continuous factor should be categorized first based on intervals with cutoff of 25, 50 and 75 percentiles before they could be used as stratification factors; categorical factors however could be used directly.

3.6 Missing Data

For each factor, if more than 30% of all observations were missing, then the factor should be removed from the model since no imputation method could supplement the lost information without enough observed data; if more than 15% of all observations from a

factor were missing, then the missing values from the factor would be imputed. For continuous factor, imputation was carried out using the least squares multiple regression incorporating optimum transformations; and for categorical variables, missing values were imputed using recursive partitioning.

3.7 Variable Reduction

Completely independent factors do not exist in medical data; however including correlated factors in the model may lead to incorrect inferences and parameter estimates. To avoid harmful multicollinearity, variable reduction or clustering was intended. The reduction procedure was carried out using the original scale of the factors instead of the transformed values, so that the patterns were only derived from the original factors that were actually related.

Variable reduction or cluster analyses were executed through unsupervised learnings[144] to determine clusters; a hierarchical cluster analysis was used to determine the linkage among factors; once clusters were confirmed, composite factors could be derived through a principal component analysis unless there was other better alternatives. Once the composite factors were derived, they were included in the survival model in place of the original factors (see section 3.7.3 for details) for further analysis.

3.7.1 Multicollinearity

Spearman's ρ^2 rank correlations were calculated to assess the multicollinearity among covariates. The following analysis was performed to remove multicollinearity.

3.7.2 Principal Component Analysis (PCA)

For categorical factors, the correspondence analysis (see section 2.5 for a brief description of the approach) should be useful for factor reduction, however for a mixture of categorical and continuous factors, the principal component analysis (PCA) with total maximum variance[128] method should be used to convert all factors into linearly uncorrelated principal components. Only the factors that were determined to be clustered together should be processed via principal component analysis, once principle components were constructed, they should be used as the composite factors to replace the original clustered factors for further analysis, with the principal component score as the values of the composite factors.

3.7.3 Variable Reduction or Cluster Analysis

Relationship among factors was assessed through hierarchical cluster analysis with Hoeffding's *D* statistic; clusters can be determined using the pedigree map among factors as obtained from the hierarchical cluster analysis. Only clusters picked up by hierarchical cluster analysis were used to derive the composite factors.

To derive the values of the composite factor from the original factors, reinforcement algorithm should be followed as the first choice, if there was one from past experience or from external experts. However, when such algorithm did not exist, the algorithm for deriving composite factor should be achieved by obtaining the predictive cluster score from a subset of all factors in this cluster (using linear regression or recursive partition). The subset of factors were obtained from the tentative fit of the survival model fit; only the factors in the cluster were fit to the survival data using a tentative Cox regression model with the 1st order linear form of all factors in that particular subset (since this was a tentative fit, no need to check for functional forms, interactions or proportionality), and run a model selection, the subset of the factors that remains in the final model should be the ones with predictive power. Thus this subset of the factors should then be processed with the principal component analysis as discussed in section 3.7.2 to derive the values of the composite factor (or scores for the components).

3.8 Description of Graphic and Non-Graphic Tools

3.8.1 Normality

Normality was checking using the following tests, including D's Agostino's K-squared test [10], Anderson-Darling test[11], Kolmogorov-Smirov test[145, 146] and Shapiro-Wilk Test[13]. Density plots were also used to confirm the normality test; if normality assumption was significantly violated, factors had to be transformed to ensure normality.

3.8.2 Transformation Plot

Each transformed factor was plotted against the original factor to determine proper transformations. If factors were already normal, transformations should not be needed for most of the models. However, for complicated models, such as penalized and partial least squares models, normalization may be needed to achieve convergence efficiently; yet for these models, normalization was part of the process when the model was implemented,

which should be automatically carried out within the modelling algorithm internally, and once complete, the model should be able to transform the normalized factors back to the original scale. Therefore, external normalization should not be needed unless departure from normality was detected.

3.8.3 Nonlinearity and Heteroscedasticity

For categorical variables with more than 2 categories, typical Cox models should create dummy variables internally to represent the pairwise difference between categories; however in the two studies it was more convenient to create dummy variables externally to represent the relative difference between any two levels of the categorical factor; either way categorical factors should not have nonlinearity concern, since categorization was nonparametric in nature. For continuous factors, transformation using nonlinear functional forms was a parametric process, it should be crucial to identify the proper functional forms for continuous factors; while graphic display of the Martingale residuals was a reliable and effective tool. Therefore, whenever possible, it was advisable to obtain the Martingale residuals from the Cox PH linear model including all continuous factors in their original linear forms as the only covariates, after which, the residuals were plotted against the factors in their original form to reveal the functional form for further transformation. The same plot could also be used to detect heteroscedasticity, if significant heteroscedasticity was detected, robust standard error should be used for inferences instead. However, if residual plots were not always available, especially when there were too many factors to be considered, then Wald test on the deviance difference between Cox models with and without the nonlinear forms could be used to narrow down the potential nonlinearities.

If nonlinearity was confirmed, *RCS* and *FP* transformations would be intended for the typical Cox regression models. This step was not needed for the nonparametric RSF, penalized Cox regression or derived input Cox regression models (PCR or PLS Cox models). Details of the *RCS* and *FP* transformations are presented below. Given a nonlinear factor, x_l:

- Restricted cubic spline (*RCS*)[43]: the minimum number of knots (u) for the *RCS* was the number of inflection points from the residual plot on the continuous factor + 2:

$$f_l(x_l) = \alpha_0 + \alpha_1 x_l + \sum_{h=2}^{u} \alpha_h (x_l - t_h)^3 \quad \text{..... Eq. 48}$$

where t_h's were usually selected using the corresponding quantiles of x_l. Note that for each *RCS* transformation, without considering the intercept, a total of $u + 1$ regression coefficients had to be estimated, but only need to spend an extra of $u - 1$ dfs comparing to the linear form, because the coefficient α_{h+1} was derived from α_h and α_{h-1}.

- Multivariate fractional polynomial (MFP)[147] transformation, the 4-df fractional polynomials (*FP2*) with power p_1, p_2 $(m = 2)$ from a pragmatically chosen restricted set $V = \{-2, -1, -0.5,\ 0,\ 0.5,\ 1,\ 2,\ 3\}$, where x^0 denote $\log x$. The nonlinear form of x_l, is

$$f_l(x_l) = \sum_{h=1}^{m} \alpha_h x_l^{p_h} \quad \text{..... Eq. 49}$$

FP1 of 1-df: $m = 1, \alpha_1 \neq 0, p_1 = 1$; 2-df *FP2*: $m = 2, \alpha_1 \neq 0, p_1 \neq 1$; 3-df *FP*2: $m = 2, \alpha_1 = \alpha_2, p_1 \neq$; 4-df *FP*4:$m = 2, \alpha_1 \neq \alpha_2, p_1 \neq p_2 \neq 1$, it become a 4-df *FP* test.

3.8.4 Interactions

For identification of interactions, Martingale residual plot was also an effective tool to examine the possibilities. For a pair of continuous factors, any one factor from the pair were categorized into 4 intervals using the 25, 50 and 75 percentiles of the same factor; the residuals were plotted against the other continuous factor in that particular pair stratified by the categorized factor. For interaction between continuous and categorical factors, interaction was visually checked by Martingale residual plot against the continuous factor stratified by the categorical factor. A separate smooth spline curve was fitted for each level of the stratification factor; interactions should be confirmed if lines were not parallel due to different slopes. And once the interaction between every pair of factors was confirmed, all interaction terms between the functional forms of the two factors were included in the model. For a pair of categorical factors, interactions were evaluated using Deviance Wald test, because the graphical display was not very intuitive. Additionally, when it was not possible to get the Martingale residual plots or it was visually impossible to examine the graphic displays of all possible interactions, the

deviance difference between the two models with and without the interaction terms could be used to screen the potential interaction terms.

3.8.5 Proportionality and Time-Dependent (Varying) Covariate(s)

To verify proportionality assumption, typical Cox PH model was tested against all terms, including the nonlinear functional forms of all factors and interaction terms as identified from section 3.8.3 and section 3.8.4.

Proportionality assumption was checked using a global Chi-square test[15] of the scaled Schoenfeld residuals[14] on time (or a function of time, if transformation of time was deemed necessary), a p-value of <0.05 would suggest significant non-proportionality; to be more conservative, a p-value of $\lesssim 0.10$ would suggest clues for non-proportionality. However, since this test was neither sensitive nor stable, a scaled Schoenfeld partial residuals were obtained from Cox regression, the residuals were plotted against each term (including the functional forms of the factors and the interaction terms) to verify the proportionality assumption.

If non-proportionality was confirmed, extension of Cox PH model should be considered, in order to relax or fix the non-proportionality; then it is necessary to discuss the slight differences between the time-varying factors and time-dependent factors. Time-varying factors refer to factors that change over the life time of subject; time-dependent factors refer to the factors that do not change over time but confounded with survival time. The difference is very subtle, but the solutions are slightly different. Time-varying factor can be modeled using both extension 1 and extension 2 as described below; in fact, the two extensions should applied in sequence, option 1 is the first choice for adjustment of time-varying factor; if option 1 does not completely resolve non-proportionality, option 2 can be applied on top of option 1. However, the time-dependent factor should be adjusted using extension option 2.

1. Andersen-Gill extension of the Cox PH model (see section 2.3.3.3 for the formulation of the model) should be implemented for adjustment of time-varying factor; the time-varying factor was considered independent of the failure event, therefore the subject who had time-varying factor was assumed to have multiple observations, each with a single factor, censoring time (the time when the first varying was detected) or failure event time.

2. Time-dependent extension of Cox PH model should include a second order interaction term(s) (see section 2.3.3.2 for details); the interaction term(s) should be constructed between $g(t)$ (transformation function of time) and the actual functional form of the factor(s) that depart from proportionality. The actual transformation function for time was determined solely based on experience. Therefore it is possible that the non-proportionality may not be completely fixed; then other alternatives may have to be intended (i.e. using the time-dependent factors as stratifications).

Once non-proportionality was fixed or proportionality assumption was relaxed, different approaches based on multivariate Cox regression model with nonlinear transformations could then be performed following the procedures from section 3.3.

3.9 Training, CV and Testing

From non-statistician's direction, it is always sound to partition the original survival data into training and testing set, of which, training is carried out on the training set for bias reduction and testing is carried out using the test set for assessing the variance or prediction performance. However, statisticians would like to use the entire dataset for model selections; since it is considered a waste of information to use only a portion of the original data for statistical analysis. The topic has been quite controversy, therefore in this research, a compromise between the two was considered: the original survival data was randomly partitioned with 3:1 ratio, 75% of the original survival data was used for training and 25% was used for testing; whenever it was possible as long as the intended model was estimable, i.e., the training set was large enough to estimates the coefficients for all covariate, the model selection should be carried out using the training set (except for the RSF), a 10-fold leave-one-out CV (see section 3.10 for details) should be performed over the training set for cross validation (CV) performance of the model; prediction of future events should be based on the test set and prediction performance should be evaluated over the test set similarly. However, if the intended model did not converge with the training set due to deficiency of data, in other words, if the training set did not have enough data points to estimate the coefficients for all covariates, then the entire survival data should be used instead. For the nonparametric random survival forest (RSF), the approach was based on bootstrap aggregation, which was used to select the

model corresponding to the best CV performance, therefore no separate CV was needed; otherwise it should follow the same procedure.

For typical Cox regression models, CV should be carried out using model AIC based on the training set and the best model should be selected to achieve minimum AICs of all models. For penalized Cox models, the CV performance could be measured using partial log likelihood deviance or cross validation errors. Cross validation was carried out using the following steps:

Randomly partition the survival data into 10 equal-sized subsamples.

Do h=1 to 9;

Hold out subsample *h;*

Train the model based on the remaining 90% of the data;

Use subsample h, to validate the model selected from the training set;

End;

Calculate the average cross validation error across the 10 hold-out sets of predictions

$$\mathrm{CV}(\hat{f}) = \frac{1}{10}\sum_{h=1}^{10} L(y^{(h)}, \hat{f}^{-\kappa(h)}\left(x^{(h)}\right) \quad \text{..... Eq. 50}$$

The prediction performances were measured with the test set; detailed specifications will be discussed in section 3.11. For penalized Cox regression models, the partial log likelihood deviance was used for cross validation. It was just -2 log(likelihood), the minimum of partial log likelihood deviance should correspond to the maximum of the partial log likelihood; the model with the minimum of the partial log likelihood deviance should have reached its best performance.

Additionally, the selected Cox models should be validated using the following statistics, such as Somers' D_{xy} rank correlations¸ index of unreliability (U), discrimination index (D), overall quality (Q), slope of the overall calibration, the maximum absolute difference in the calibrated probabilities (E_{max}), concordance index (c-index). The Somers' D_{xy} is a rank correlation between the predicted and the observed survival status. The discrimination index, D, is measured as the model likelihood ratio $(\chi^2 - 1)/n$; it is used to measure the ability to distinguish the two survival status. The index of unreliability (U) is the difference in -2 × log likelihood between $X\hat{\beta}^{(u)}$ with the uncalibrated slope, $\hat{\beta}^{(u)}$, from the training sample and $X\hat{\beta}^{c}$ with the calibrated slope, $\hat{\beta}^{c}$,

from the test sample divided by n. The overall quality index is measured as $Q = D - U$. The slope of the overall calibration is the lowess curve of the predictions vs. the actual survival status. B-index is the Brier and quadratic probability score. The g-index is Gini's mean difference of each pair of predicted survival vs. observed survival status, weighted by the regression coefficients of the covariate terms including the original factors. The c-index is an overall measurement of concordance and discordance.

3.10 Analysis Procedures

The overall procedures for analysis are listed below.

1. Prepare data for analysis;
    A. Normality tests were performed; density plots for all factors were prepared for confirmation of normality. Transformation was need if departure was detected.
    B. Missing values were imputed as needed (see section 3.6).
    C. Nelson-Aalen estimates or KM estimates of the survival probability should be obtained without consideration of any covariates or stratification factors; this was to confirm legitimate application of Cox model.
    D. Variable reduction or cluster(see section 3.7 for details): perform multicollinearity analysis to identify highly correlated factors; use hierarchical cluster analysis to detect the linkage among factors and to identify potential clusters; fit a typical Cox PH model to select a subset of the factors within each potential cluster; construct the composite factor with existing reinforcement algorithm, or perform PCA (see section 3.7.2 for details) to construct principal component for the cluster of factors, component scores from the PCA should be used as the values for the new composite factor.
    E. The survival data should be randomly partitioned into training (75%) and testing (25%); for survival data with multiple records per subject, extra caution should be taken in order to retain the within subject correlation structure, the partition should be based on subject identifiers, if a subject is selected for the training set, all records from the same subject will be stay in the training set; the same should be followed for test set.

F. Use the training set to determine the initial survival model for model selection:
    i. In general, model assumptions should be carefully checked, proper functional forms of factors should be identified, interactions had to be checked and proportionality assumption had to be confirmed (time-varying effect had to be properly adjusted and non-proportionality had to be fixed if violation of proportionality assumption was detected); CVs were attempted for model selections; test set was used for prediction future outcomes and for assessing prediction performance of the selected model. However, it has been controversy about how to fully utilize the original data, therefore in some special cases, the original data may be used instead of the training set due to deficiency of data points; for example, in the real world case study, the original NKI70 data was used to train the PCR model, since the PCR model was unable to converge with the training set.
    ii. Conventional Cox regression model was fit to the training set (see section 3.1 for details).
    iii. Multivariate Cox regression model with *RCS* transformation was fit to the training set (see section 3.2 for details).
    iv. For *FP* transformation:
    Perform a 4 df test at the level of the best-fitting second-degree *FP* against the null model. The model should automatically identify the best transformation for each factor, then Cox model with the best *FP* transformation was fit to the training set (see section 3.2 for details).

2. Model selection or CV
    A. Model selection was performed for the typical Cox PH Regression with AIC as the selection criteria.
    B. Random survival forest was (see section 3.3.1.2 for details) were cross validated within the bootstrap aggregation algorithm ;

C. Penalized Cox regression (see section 3.3.2.1 for details) models were cross validated with partial log likelihood deviance or CV errors whichever was appropriate.
D. Partial least squares regression (see section 3.3.3.2 for details) was tuned with the training set.

For analysis of different datasets, there should be subtle differences in these procedures, but the general framework should remain the same: training (and/or CV) and testing. The sequence of models for each approach were trained using the training set and the best model was selected based on AIC, partial log likelihood deviance or mean CV error as appropriately for the survival model; prediction of future outcomes and prediction performance should be evaluated over the test set; the prediction performance were then compared across different models. As a general rule, the best model was selected if the most improvement were achieved with the minimum number of covariates with the only exception of the nonparametric random survival forest (RSF) approaches, where a separate cross validation step was not needed CV was done within the bootstrap aggregation.

3. Predict future or unseen outcomes based on the best model with the selected covariates from step 2; estimate prediction performance of the best model based on the test set and bootstrap the test set to obtain the 95 percentile credible intervals (PCI) and compare the prediction performance across all survival models.

3.11 Evaluation of Prediction Performance

As discussed previously, several different measurements or statistics were developed to assess the model performance in this research, Brier score, AIC, partial log likelihood deviance and time-dependent AUCs. The first three statistics are considered as a loss function, the best model is the one to minimize the loss function; the last one is for prediction powers, the best model should achieve the best prediction AUCs.

For typical Cox regression models, Akaike information criterion (AIC) was used to select the best model via CV; for penalized Cox regression models, one option was to use partial log likelihood deviance as the selection rule via CV, however it was found that

Brier score was more reliable for model selection; additionally, nonparametric RSF modes were also cross validated using Brier scores, which was named as out-of-bag errors (OOB).

For test set, the time-dependent prediction error as measured by Brier scores was one of the statistics for assessing the prediction accuracy of all survival models; a perfect survival model should have zero prediction errors. The time-dependent AUC was another measurement for assessing the prediction performance of the models; theoretically, a perfect survival model should achieve 100% AUCs, but in practice, survival models can never reach 100% AUCs, instead a survival model with ≥65% prediction AUCs would be a reasonable model, ≥70% AUCs would be good, ≥75% would be exceptional, and ≥80% would be excellent; a prediction AUC of 50% should be no better than a random guess. The formula for Brier Score, AIC and partial log likelihood deviance are:

Brier Score at t $$L(t) = 1/N \sum_{j=1}^{N} \left(Y(t) - \hat{Y}(t)\right)^2 \quad \text{.....Eq. 51}$$

AIC:	$\mathrm{AIC} = 2\nu - 2\log(\mathrm{Likelihood})$ .....Eq. 52
Partial Log Likelihood Deviance	$-2\log(\mathrm{Likelihood})$

where ν is the number of covariates in the model; $Y(t)$ is the observed survival status at time t, which is $I(T_j \geq t)$, $\hat{Y}(t)$ is the predicted survival probability at time t, N is the total number of subjects. The prediction error, L, is used to estimate the prediction performance at time t, which is also known as the Brier score[148] at t. The integrated Brier score is obtained by averaging the total Brier scores for interval $(0, t]$, which can be estimated with $1/k \sum_i BS(t)$, where $i = 1 \ldots k$ different event times. For survival models, the empirical time-dependent Brier score $BS(t)$ can also be obtained as

$$\mathrm{BS(t)} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left\{ \frac{\hat{\mathrm{S}}(\mathrm{t}|\mathrm{x_i})^2 \mathrm{I}(\mathrm{t_i} \leq \mathrm{t} \wedge \delta_\mathrm{i} = 1)}{\widehat{\mathrm{G}}(\mathrm{t_i})} + \frac{\left[1 - \hat{\mathrm{S}}(\mathrm{t}|\mathrm{x_i})\right]^2 \mathrm{I}(\mathrm{t_i} > \mathrm{t} \wedge \delta_\mathrm{i} = 0)}{\widehat{\mathrm{G}}(\mathrm{t})} \right\} \quad \text{... Eq. 53}$$

with individual survival time t_i, censoring indication δ_i, and the estimated survival probability $\hat{S}(t|x_i)$ at time t based on the prognostic model given covariate values x_i for subject i out of n subjects (Graf et al. 1999). $\widehat{\mathrm{G}}(\mathrm{t})$ denotes the Kaplan-Meier estimate of the censoring distribution at time t, which is based on observations $(t_i, 1 - \delta_i), i = 1, \ldots, n$. I stands for the indicator function.

The time-dependent AUC measurement is were very much related to the c-index, as recommended by Harrell (2012), which is the probability of concordant pairs of predicted and observed survival status among all pairs of responses. Specifically, if the predicted survival probability is larger for the subject who (actually) lived longer, the predictions for that pair are said to be concordant with the (actual) outcomes. The c-index is an overall measurement of the prediction performance. But similar to the time-dependent prediction errors, the probability of concordance at each time point can also be evaluated by assigning each subject a survival status at the given time point: if the time point of interest occurs prior to the event time for a specific subject, the subject should be alive (status of 0) at that particular time point; if the time point of interest occurs after the subject had an event, the subject should have already been censoring (status of 1) at the particular time point; otherwise, the subject should have status of 0 at the time point. Consequently, at each time point, the predicted survival probability of each subject from the test set could be checked against the survival status of the same subject at the corresponding time point; averaging all subjects in the test set at the same time point, the probability of all concordant pairs could be obtained for the given time point; then the probabilities of concordance at different time points are referred to as time-dependent AUCs, it can be estimated as

$$AUC(t) = Pr\{\hat{S}_i < \hat{S}_j | T_i < T_j,\ T_i \le t\} = \frac{\sum_{T_i \le t} I(\hat{S}_i < \hat{S}_j) \times I(T_i < T_j)}{\sum_{T_i \le t} I(T_i < T_j)} \quad \text{..... Eq. 54}$$

where $\hat{S}_i$ and $\hat{S}_j$ are the predicted survival probability for subject i and j at time t, $i, j = 1, 2, \ldots, n$ and $i \neq j$.

For prediction errors and time-dependent AUC measurements, the robust estimates and the corresponding 95 percentile credible intervals (PCI) were obtained based on the mean and the 95 percentiles of the measurements from 1000 bootstrapped samples of the test set, unless noted otherwise.

3.12 Software Packages

All data were processed using SAS/BASE Software, Version 9.2 of SAS[149] System for Windows (2011, SAS Institute, Cary NC) and analyses were performed using R[150] Software for Windows, Version 3.03 or higher (R Core Team).

Chapter 4. Simulation and Case Studies

One simulation and one real world case study were used to evaluate different survival models in this research paper. The data from the simulation study was simulated with a computer model, with which a time-varying treatment effect was also generated. The real world case study was performed on a breast cancer dataset downloaded from Netherlands Cancer Institute for evaluation of metastasis-free survival[143]; in this dataset, a total 5 clinical factors and 70 gene signature profiles were collected from 144 subjects.

4.1 Simulation Studies with Time-Varying Treatment Effect

The survival data for the simulation study was generated from a computer model (see section 4.1.1 for detailed descriptions); once complete, data was randomly partitioned into training and testing set in 3:1 ratio. As previously mentioned in section 3.9, all survival models were initially trained with the training set, 10-fold cross validation was carried out over the training set to select the best subset of factors and the best survival model, the test set was used to predict future survival outcomes and to evaluate the prediction performance. In the simulation study, a time-varying treatment effect was also generated, a subject could have received different treatments at different time during the study, i.e., some subjects may have more than 1 observation, each with a different treatment. In order to retain the time-dependency and within subject correlation, the partition of training and testing sets was carried out on subject-level; each subject in the study was assigned a subject identifier; if a subject was selected for the training set, all of the observations from the same subject went into the training set; the data from the rest of the subjects comprised the test set.

4.1.1 Description of the Survival Data Simulation

A total of 2000 subjects were simulated using R software; 7 factors included age, sex, race, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI) and treatment were generated from computer programs. The factors, SBP and DBP were highly correlated (correlation coefficient of 0.9) and either was weakly correlated with BMI (the correlation coefficient between SBP and BMI was 0.3 and the one between SBP and BMI was 0.2). Once generated, the factors were used to generate

the survival data following computer model. After which, a time-varying treatment factor was added based on a simplified but typical scenario, subjects were randomized to active and placebo treatment at the time of enrollment (the placebo treatment included active control such as standard of care (SOC), inert tablets or sham treatment). In this simulation study, the SOC was assumed for the placebo treatment. Subject were allowed to switch treatment after the 1 Year post-randomization; for simplicity reasons, only placebo treated subjects were programmed to switched to the active treatment after 1-year post randomization, which was quite common in phase III, IV and epidemiology studies. In a typical clinical study, subjects are usually closely monitored and are not allowed to switch treatment during the efficacy evaluation period (such as 1 year for this simulation study), after subjects reached the target duration, they should be allowed to switch treatment or therapeutic options per investigator's discretion, required by the study protocol (to enter into an open-label phase) or any reasons other than random decisions. The same scenario was followed to simulate the time-varying treatment effect. The reasons for switching treatment could be very subjective, but the subjectiveness could not be simulated without complicated computer models, which was not the primary interest of this research. Thus, a simplified computer model was used to simulate the treatment switching (see below for details); the model was much more simplistic than any typical real world clinical trials, but the analysis approaches and models should be general to all studies.

- Age was randomly generated from a normal distribution with a mean of 50 and a standard deviation of 12 years; Sex was a binary variable, randomly sampled from the set of {Male, Female} with replacement based on 6:4 ratio; Race was an ordinal variable, which was randomly sampled from the set of {White, Black, Hispanic and Asian} with replacement using 55:23:16:6 ratio, 55% of White, 23% of Black, 16% of Hispanic and 6% of Asian;
- SBP, DBP and BMI were generated from random normal distributions with correlation matrix of $\begin{bmatrix} 1 & 0.9 & 0.3 \\ 0.9 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix}$ and with mean (SBP) = 110 and SD (SBP) = 9 (mmHg), mean (DBP) = 75 and SD (DBP)= 5 (mmHg), mean (BMI) = 28.5 and SD (BMI) = 4 (kg/m^2).

- Subject were randomized to active treatment or placebo (SOC) at the beginning of the study based on 2:1 ratio stratified by subject sex; at 1-year post randomization, placebo treated subjects were allowed to switch to active treatment (23.9% actually switched), the subjects from the active treatment group continued their treatment until the end of the study. In theory, subjects from the active treatment group could also switch to placebo or other treatment options, which could be programmed using the same model as we did for placebo treated subject, but it should just add more complexity to the simulations, nothing more. Furthermore, in a real world setting, subjects could switch back and force between treatment groups until they were censored or failed. However since the simulation model was not the focus of the research, it was not worth the effort to simulate the extra scenarios, such as subjects switching from the active to placebo treatment or switching back and force, since the same survival model(s) should be able to account for all these scenarios. Thus for the sake of simplicity, the extra scenarios was not simulated.
- The hazard function was simulated using an exponential model. At baseline, male and female subjects had different hazard functions; subjects under different treatment should also have different hazard (subjects under active treatment were assumed to have smaller hazard). While subjects randomized to placebo at baseline were allowed to switch to active treatment after 1-year post randomization; they should follow the hazard function for active treatment starting at the time of switching, depending on their sex and the actual treatment received. But for subjects who stayed in the same treatment, their hazard function should remain the same.

Female $\lambda(t|X)) = .02 \times \exp\{.16 \times \sqrt{\text{age}} + .8 \times \{\text{Female}\} + 0.07 \times \{\text{White}\} + 0.8 \times \{\text{Black}\} + 0.08 \times \{\text{Hispanic}\} - 0.6 \times \{\text{Active Treatment}\} + .1 \times [(\text{MAP} - 91)/5.96]^3\}$ ..... Eq 55

Male $= .02 \times \exp\{.16 \times \sqrt{\text{age}} + .8 \times \{\text{Female}\} + 0.07 \times \{\text{White}\} + 0.8 \times \{\text{Black}\} + 0.08 \times \{\text{Hispanic}\} - 0.6 \times \{\text{Active Treatment}\} + .6 \times [(\text{MAP} - 91)/5.96]^3\}$

- Anticipated time of follow-up or drop out (censoring time), ctime, and proposed time of event (event time), etime, were generated from, $\text{ctime} \sim \text{Uniform}\ (0, 14)$ and $\exp(-\lambda^{(1)} \times \text{etime}^{(1)}) \sim \text{Uniform}\ (0, 1)$ for subject who did not switch treatment arm until the end of the study;
  - For subjects who stayed in the same treatment until the end of the study: if etime ≤ ctime, the subject had an event at etime; otherwise censoring occurred at ctime;

    After 1-Year post baseline, placebo treated subjects who were at risk at 1-Year were allowed to switch to active treatment with a probability of 60%: stime ~ Uniform (0, 14); upon switching to active treatment, these subjects should follow the hazard function from the active treatment, then $\exp(-\lambda^{(2)} \times \text{etime}^{(2)}) \sim \text{Uniform}\ (0, 1)$.
  - For those subjects who switched treatment, if $1 + \text{stime} + \text{etime}^{(2)} \leq \text{ctime}$, event occurred at the time of $\text{stime} + \text{etime}^{(2)}$; otherwise it was censored at ctime. Here the superscript $^{(1)}$ or $^{(2)}$ is used to refer to the different treatment phase before or after switching to the different treatment, respectively, for the subjects who switched treatment.

4.1.2 Results of the Simulation Study

4.1.2.1 Summary Statistics of the Simulation Study

The demographics of the study population are summarized in Table 1; as can be seen from the table that all factors were equally distributed between the two randomization groups. A summary of subject survival status at 1-year post baseline and treatment switching after 1-year post baseline is presented in Table 2. For this simulation study, a total of 106 (10.6%) subjects from active treatment arm had events prior to 1 year and 163 (16.3%) subjects from the placebo (SOC) arm had events prior to 1 year. Of the 891 subjects randomized to active treatment who survived 1 year, none of the subjects switched to placebo; of the 840 subjects randomized to placebo and survived 1 year, 201 (23.9%) switched to the active treatment, and the rest of the 639 subjects stayed in the placebo (SOC) arm until they censored or failed. In this study, the treatment switching from placebo to active was assumed to be independent of the failure event, which

however may not be reasonable in practice; in a clinical study, patients may switch treatment due to adverse events or lack of efficacy.

Table 1. Demographics of Simulation Study

Factors	Statistics	Enrollment Randomization		All Enrollment (N=2000)
		Active Treatment (N=997)	Placebo SOC (N=1003)	
Age	Mean ± SD	49.8 ± 11.73	50.5 ± 11.73	50.2 ± 11.73
	Median (25%, 75%)	49.4 (42.2, 57.9)	50.3 (42.7, 58.7)	49.7 (42.5, 58.2)
	Min – Max	16.8 – 85.11	15.8 – 94.5	15.8 – 94.5
Sex	Male	596 (60%)	623 (62%)	1219 (61%)
	Female	401 (40%)	380 (38%)	781 (39%)
Race	White	534 (54%)	546 (54%)	1080 (54%)
	Black	215 (22%)	242 (24%)	457 (23%)
	Hispanic	172 (17%)	181 (18%)	353 (18%)
	Asian	76 (8%)	24 (3%)	110 (6%)
SBP	Mean ± SD	110.6 ± 8.64	110.3 ± 8.64	110.4 ± 8.64
	Median (25%, 75%)	110.2 (104.8, 116.5)	110.5 (104.6, 115.7)	110.3 (104.6, 116.1)
	Min – Max	82.9 – 138.6	83.9 – 137.2	82.9 – 138.6
DBP	Mean ± SD	75.2 ± 4.87	75.2 ± 4.87	75.2 ± 4.87
	Median (25%, 75%)	75.2 (71.9, 78.6)	75.3 (72.0, 78.4)	75.3 (72.0, 78.5)
	Min – Max	58.4 – 89.66	58.0 – 91.5	58.0 – 91.5
BMI	Mean ± SD	28.8 ± 3.98	28.7 ± 3.98	28.7 ± 3.98
	Median (25%, 75%)	28.6 (26.0, 31.6)	28.7 (26.1, 31.3)	28.7 (26.1, 31.4)
	Min – Max	13.3 – 41.4	15.1 – 40.5	13.3 – 41.4

Table 2. Summary of Subject Survival Status at 1-Year and Treatment Switching After 1-Year Post Baseline

	Status or Treatment	Enrollment Randomization	
		Active Treatment (N=997)	Placebo / SOC (N=1003)
Survival at Year -1			
Status	Event Free	891 (89.4%)	840 (83.7%)
	Failure (Event)	106 (10.6%)	163 (16.3%)
After Year-1		891	840
	Active Treatment	891 (100%)	201 (23.9%)
	SOC	–	639 (76.1%)

NOTE: None of the subjects who were randomized to active treatment at baseline switched to placebo during the study; "–" means "NA".

4.1.2.2 Data Preparations

Before proceeding with actual analysis, the data had to be prepared for analyses, such as variable transformation, missing data imputation, normality test, variables reduction or clustering, checking for multicollinearity and heteroscedasticity, checking for linearity and determination of the correct functional form for each factor, testing on interactions, examination of proportionality assumption, etc. For this simulation study, missing values were not simulated; therefore missing data imputation was not performed. Otherwise, data preparation was carried out using the training set.

4.1.2.2.1 Normality assumption

Normality assumption was checked for all continuous factors; none of 4 normality tests (including D's Agostino's K-squared test, Anderson-Darling test, Kolmogorov-Smirov test, and Shapiro-Wilk Test) had shown apparent departure from normality for all continuous factors; the Q-Q plots (Figure 1) also confirmed the normality assumption.

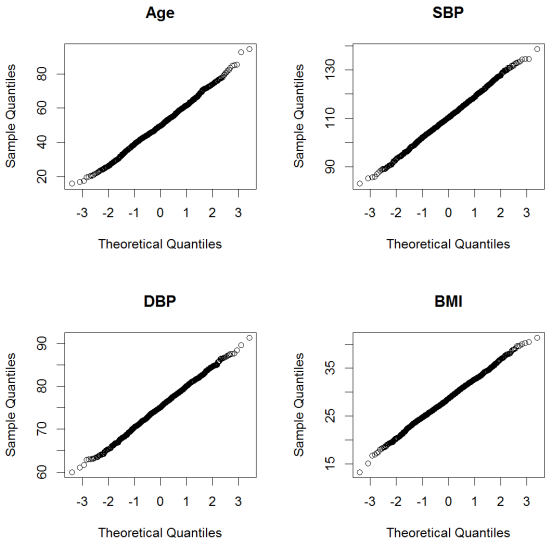


Figure 1. QQ-plot for All Factors in the Original Scale from Simulation Study

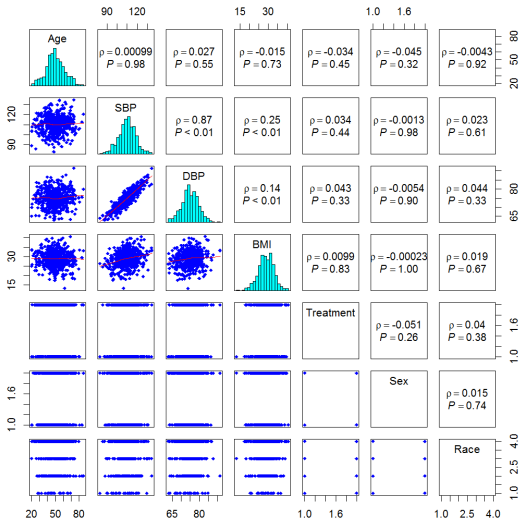


Figure 2. Scatter plot of All Factors from Simulation Study

Multicollinearity was checked using Spearman rank correlation coefficient as well as scatter plot; results are presented in Figure 2. The locally weighted scatterplot smoothing (LOESS) and the scatter plots of all factors are displayed in the lower-left triangle. Density probability histograms are displayed for all continuous factors in the diagonal of the figure (Sex and Race are categorical variables, histograms are not displayed); the

Spearman rank correlation coefficient (ρ) and the corresponding p-values (p) are presented in the upper-right triangle. It was noted that SBP and DBP had very strong correlations (ρ=0.87); and both SBP and DBP had moderate correlations with BMI (the corresponding correlation coefficients were ρ = 0.25 and 0.14, respectively). The correlations between all other continuous factors were almost negligible; the correlations involving any of the categorical variables were not very much meaningful and probably were not very informative, instead correspondence analyses should be more useful, which were not performed for this study.

4.1.2.2.2 Data Transformation/Missing Data Imputation

Data were transformed independently of the survival outcome; each transformed factor was plotted against the same factor in its original scale, the plots are displayed in Figure 3. The transformed factor of Age had a local peak at 40 and local valley around the median of 50 years, which suggested that nonlinear transformation should be needed for Age, therefore at least 3-knot *RCS* or 4-df *FP* transformations should be considered for Age; for the rest of the continuous factors, the transformed factors were pretty much linear with the original factors. Since the transformation was conducted without looking at the survival outcomes, the final decision about the transformation should be made based on the analysis results by tentatively fitting a typical Cox PH model to the data.

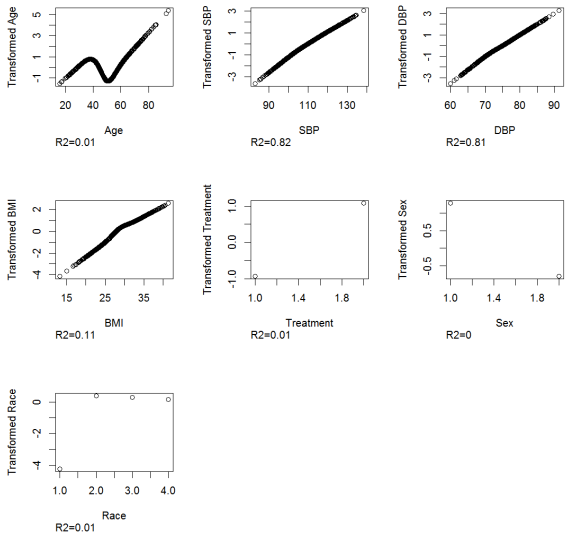


Figure 3. Transformed vs. Original Factors – Simulation Study

The survival data were simulated using a computer model; missing values were not

intentionally simulated, since the actual reasons for missingness could be very subjective and the subjective cause for missingness could not be easily simulated with a simple statistical model. On the other hand, the focus of the study was to evaluate different survival models, not on simulation models. Thus, missing values were not intentionally generated for the study; consequently, missing data imputation was not needed.

4.1.2.2.3 Variable Reduction/Clustering

A principal component analysis (PCA) was performed to transform factors into uncorrelated latent variable (components); the Sex and Race were categorical factors of character values, they had to be converted to numeric to be processed in the PCA. Table 3 presents the proportions of variance explained by each principal component without looking at the survival outcome; component 1 explained about 28% of the total variance, component 7 only explained 1% and each of the rest of the components explained about 13% to 15% of the total variance. In terms of the contribution to the total variance, component 1 was the most significant, components 2 to 6 were of equivalent importance and component 7 was probably not needed.

Table 3 Proportion of Total Variance Per Principal Components – Simulation Study

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	1.41	1.04	1.01	0.99	0.97	0.94	0.32
Prop of Variance	0.28	0.15	0.14	0.14	0.14	0.13	0.01
Cumulative Prop	0.28	0.44	0.58	0.72	0.86	0.99	1.00

The details of constitutes for each component are presented in Table 4. The 1st component consisted of SBP, DBP and BMI, contributed the most (28%) to the total variance; other than that, SBP and DBP also contributed to component 6 and 7. The categorical variables, sex, race and treatment were used as numeric variable in principal component analysis. Sex contributed to component 2, 4, 5 and 6, while race contributed to component 2, 3, 4, 5 and 6. Age however contributed to component 2, 3, 5 and 6. As can be seen easily, the two factors of SBP and DBP were always linked together for all components they contributed to; it was reasonable to combine the two variables together.

The graphic display of the "cumulative" variance of the original and transformed factors explained by each component is presented in Figure 4. For all factors in their original scale, it can be seen that the first 6 components already explained 99% of the

total variance, component 7 was unnecessary. Additionally, the figure also displayed the variance of the transformed factors explained by the components originated from the transformed factors; comparing to the variance of the original factors explained by the original components, data transformation did not add too much value, therefore variable transformations were not needed for the typical Cox regression models.

Table 4. Contribution of Factors to the Principal Components – Simulation Study

Factors	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Age		-0.566	0.558		-0.493	0.353	
SBP	-0.681					-0.119	-0.715
DBP	-0.665				-0.159	-0.217	0.694
BMI	-0.293	0.170	0.103	-0.117	0.485	0.786	
Sex (Numeric)		0.524		0.670	-0.430	0.287	
Race (Numeric)		-0.340	-0.817	-0.126	-0.298	0.333	
Treatment (Numeric)		-0.509		0.717	0.463		

Survival data were tentatively fit using the principal components Cox regression models, the Cox linear model, the Cox model with 5-knot *RCS* transformations, and the Cox model with 4-df *FP* transformation; the Cox linear model, and the Cox models with nonlinear transformations were used as reference.

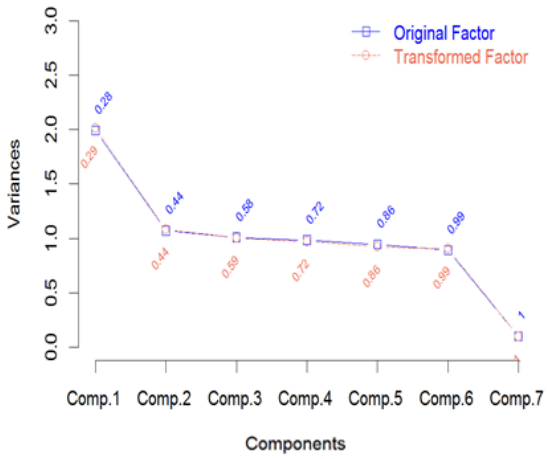


Figure 4. Cumulative Variance Explained by Principal Components (PC) – Simulation Study

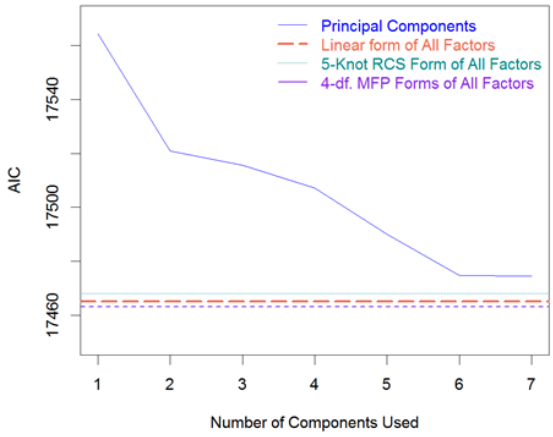


Figure 5. AIC of PCR Model, Cox Linear Model, Cox Model with *RCS* and *FP* Transformations – Simulation Study

Figure 5 displays the AIC of Cox models with one-component, two components, …, all 7 components. The reference lines were based on the AIC of 3 typical Cox models, Cox PH model with linear form (or Cox PH linear model), Cox PH model with 5-knot

RCS transformations and 4-df *FP* transformations; all 3 reference lines were well below the principal components regression (PCR) model, which indicated that PCR was probably not a good choice. Additionally, the Cox linear model and the Cox model with *FP* transformation had the similar performance and both were slightly better than the Cox model with *RCS* transformation.

Figure 5 also confirmed that component 7 was probably not needed, because AIC of the PCR model with the first 6 components already dropped to the lowest point and addition of component 7 did not improve the model performance. Moreover, comparing the PCR model with the Cox PH linear model and the Cox model with 5-knot *RCS* transformations or 4-df *FP* transformations, the PCR model had the worst AIC even with all components included in the model, but the Cox PH linear model and the Cox model with 4-df *FP* transformations were slightly better than the Cox model with 5-knot *RCS* transformations, even though the difference (in AIC) was very minimal, it seemed that nonlinear transformation of the continuous factors might not make significant improvement to the model performance.

Before further investigating the actual functional forms for all continuous factors, variable had to be clustered or reduced to eliminate multicollinearity and reduce dimensionality. A hierarchical cluster analysis based on Hoeffding's *D* statistics was performed; the linkage pedigree is displayed in Figure 6. As seen from the figure, SBP and DBP were close "siblings"; a composite variable, mean artery pressure (MAP), was constructed from the two factors, using the reinforcement learning ($\mathrm{MAP} = \mathrm{SBP}/3 + 2 \times \mathrm{DBP}/3$). As described in section 3.7.3 the reinforcement algorithm for MAP was better than the principal component from the PCA, since it was better for interpretation.

Besides the two blood pressure factors, it can be seen from the pedigree that BMI was also weakly linked to the 2 blood pressure factors, except that the linkage was too weak to be combined with the 2 factors. If however, BMI should have stronger correlations with the 2 blood pressure factors, a typical Cox PH model with the 3 factors had to be performed, a subset of the factors as selected by the model should be considered as a cluster. The cluster (subject) of factors should be processed using a principal component analysis with the cluster of factors as the only covariates; a principal component should then be constructed from the cluster of factors; the principal component could be used as

the composite factor to replace the cluster of factors for further analysis, the value of the composite factor should be set to the component score from the PCA. However, this was not needed for this simulation study.

Additionally, the 3 dummy variables, RaceHispanic, RaceBlack and RaceWhite, referred to the relative difference between Hispanic and Asian, Black and Asian, White and Asian, respectively; the reference level of Asian was not shown in the figure. The three dummy variables also had very close linkage, but since the 3 dummy variables were all coming from the same factor (Race), a composite factor for the 3 dummy variables was not needed.

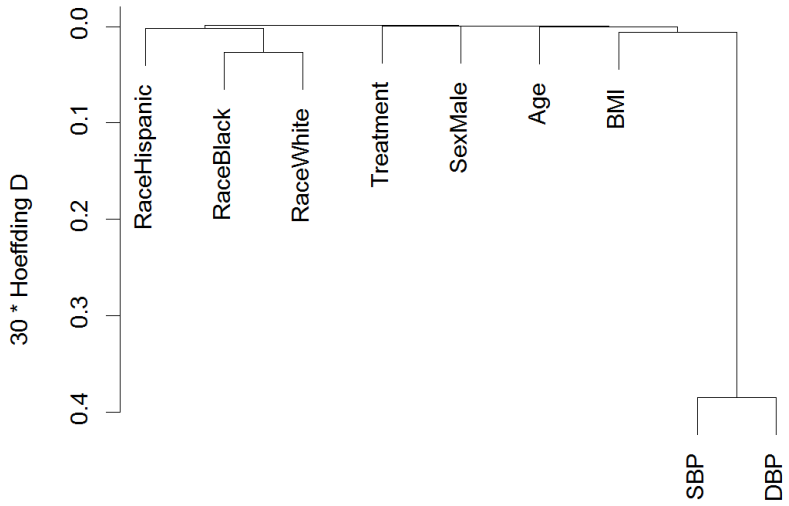


Figure 6. Hierarchical Variable Cluster Analysis – Simulation Study

4.1.2.2.4 Further Investigation of Functional Forms, Interactions and Proportionality

In this step, the initial survival model was tentatively determined, which should be used as the basis for further analysis and model selections. Potentially, there are unlimited tentative models; for example, if only polynomial terms are considered, there will be unlimited possibilities of the orders for the polynomial terms. An initial model can only account for limited possibilities; thus it should be properly declared without losing too much generality. Ideally, it is always advisable to include more terms in the initial model; since the best model is selected from the predetermined initial model, and no extra terms are added to the model for further performance improvements. Different statistical or machine learning techniques are only able to reduce the inappropriateness,

which cannot recover from an error that has already made. Then, the "best" model is the one that can achieve the best performance of all incorrect models.

On the other hand, more covariates should require more data to achieve reasonable estimates or inferences; even with the most advanced techniques, statistical inferences, predictive power, model performance, reliability and stability are also based on the size of the data; thus it is always desirable to have more data. Unfortunately, nobody can tell for sure if the size of any given data is big enough before the actual analysis. Thus a rule of thumb should be followed to predetermine the size of the data based on the number of potential factors considered (see section 3.4 for details).

Still the determination of the initial survival model is one of the most important steps and it is advisable to be more conservative for choosing the initial survival models; i.e., all possible factors should be considered for the initial model if the size of data allows; if a factor is included in the model, all potential functional forms of the factor should be included; if a factor is involved in an interaction, all functional forms of the factor should be included in the interaction.

The initial survival model should be determined only after the correct functional forms of the factors are identified, appropriate interactions are determined and proportionality assumptions are confirmed; if however non-proportionality is detected, time-dependent (or time-varying) adjustment should be considered. In this study, two approaches (*RCS* and *FP* transformations) were considered for nonlinear functional forms for typical Cox PH models:; for interactions, only second order piecewise interaction terms among all covariates (including all functional forms of continuous factors and categorical factors) were considered.

For the simulation study, all subjects were assumed to stay their original randomization group, even though some of them switched to a different treatment after 1-year follow-up following the intent-to-treat (ITT) analysis (the first analysis principal for superiority test); the plot of corresponding KM estimates of the survival curves is displayed in Figure 7. In this analysis, subjects who switched to the active treatment should theoretically follow the hazard from the active treatment arm immediately after the switching (assuming no carry-over effect), because of the treatment benefit from the active treatment, neglecting the switching should have incorrectly forced the benefit

received from the active treatment into the placebo, thus underestimate the hazard from the placebo group, which was the reason for underestimating the treatment benefit between the two groups. Moreover, the proportionality assumption was also significant violated (the scaled Schoenfeld residual of the randomization was significantly correlated with time, $\rho = -0.182$ and $P\text{-}value \ll 0.0001$); the downwarding slope of the scaled Schoenfeld residual plot as shown in Figure 8 also confirmed the non-proportionality finding.

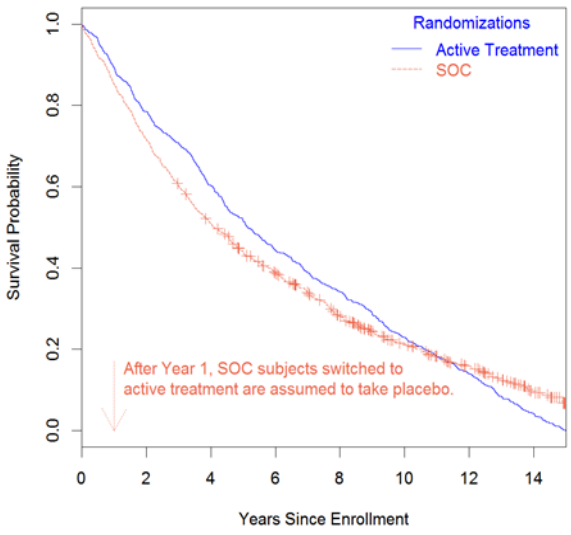


Figure 7. KM Estimates by Baseline Randomization: Subjects were Kept in their Original Randomization Group – Simulation Study

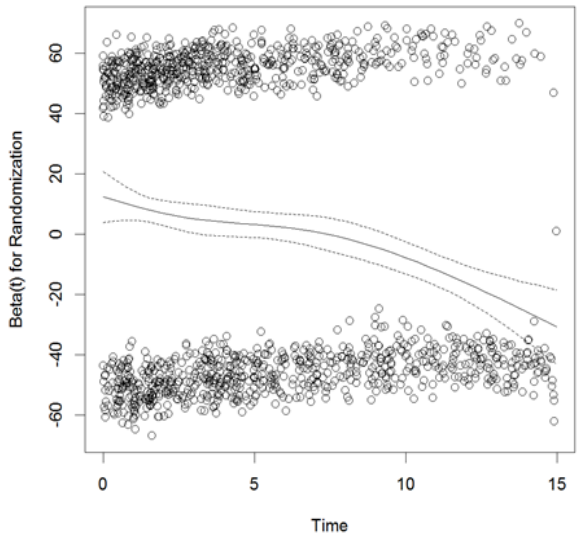


Figure 8. Scaled Schoenfeld Residuals for Treatment vs. Time: Subjects were Kept in their Original Randomization Group – Simulation Study

Another approach assumed that subjects who switched to a different treatment were censored at the time of treatment switch and the KM plot of the survival curves is not presented, since the plot is very similar to Figure 7. This approach violated the non-informative censoring assumption (previously discussed in section 1.4, section 2.3.3.4). Both of the above two approaches were significantly biased in favor of placebo and violated the proportionality assumption.

Apparently, placebo treated subjects who switched to active treatment should have received different treatments during different study period; these subjects were treated with placebo before the switch and treated with active treatment after the switch. Therefore, subjects who switched to a different treatment group should have two occurrences corresponding to the different treatment phases (only placebo subjects who

did not have event during the first treatment period were able to switch to the active treatment); thus the survival data was modeled the same way as if it was a multiple occurrence survival model, since the placebo subjects undergone treatment switching should have two observations, each with different treatment, but only one failure event occurred after the treatment switching and the treatment switching was considered as a different type of event (competing risk model).

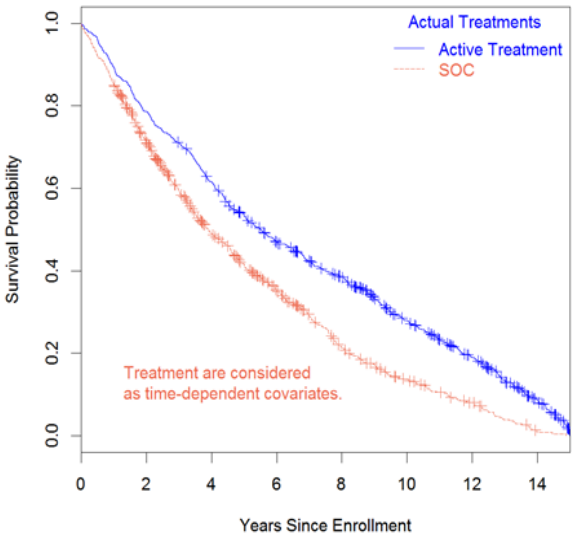


Figure 9. Cox PH Estimates of Survival Curves by Actual Treatment with Adjustment of Time Varying Treatment Effect – Simulation Study

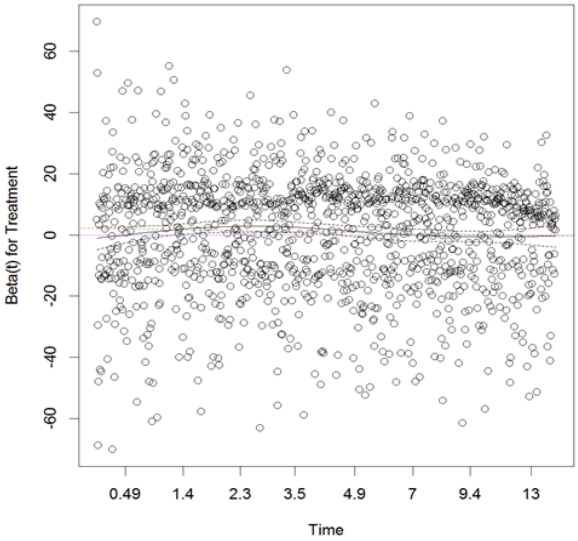


Figure 10. Scaled Schoenfeld Residuals for Treatment vs. Time with Adjustment of Time-Varying Treatment Effect – Simulation Study

To account for multiple observations per subject with consideration of time-varying treatment effect, the two extensions as described in section 3.8.5 could be implemented following the same order as described. The Andersen-Gill extension (extension option 1) of Cox PH model was first applied. The survival curves stratified by the treatment received are displayed in Figure 9 and the corresponding scaled Schoenfeld residual plot against time is displayed Figure 10. The survival curves between the two treatments did not cross in the middle until they reached Year-15 when the last failure event occurred from the last subject; this was a theoretical scenario (all subjects failed before the end of the study), which was be very rare in practice, since almost all clinical trials or studies should have already had terminated before all subjects experienced a failure event. The scaled Schoenfeld residual plot is almost parallel to the x-axis, the proportionality test was no longer significant (*p-value* ≈ 0.41). Thus non-proportionality had been properly

resolved using the Andersen-Gill extension of Cox PH model. Comparing this approach with the previous two approaches, this model should be much more reasonable.

Therefore, for the 3 typical Cox models as described in section 3.1 and 3.2, Andersen-Gill counting process extension of Cox regression models should be implemented first to adjust for the time-varying treatment effect (extension option 1 from section 3.8.5); if non-proportionality was an issue for any of the factors (including the time-varying treatment effect), extension option 2 from section 3.8.5 should be intended. The two extensions are still referred to as Cox PH models or Cox regression models.

4.1.2.2.4.1 Functional Form and Interactions

To investigate the actual functional forms for each continuous factor, *RCS* and *FP* transformations were intended; the former one was carried out using AG extension of Cox model with *RCS* transformation and the latter was carried out using AG extension of Cox model with *FP* transformations.

For Cox model with *FP* transformation, the first step was to seek the nonlinear factors using a MFP procedure, the procedure was used to estimate the power terms for each factor (see Eq. 48 from section 3.8.3 for details). At this stage, only individual factors with 4-df *FP* transformations were included in the model without considering interactions or time-dependent covariates. The MFP procedure was used to search for nonlinear factors such that the log relative hazard should have a linear relationship with the entire functional form of the nonlinear factors. As discussed in 3.8.3, the nonlinear transformations were only considered for continuous factors; categorical factors were included just for reliability check, since categorization is a nonparametric process, all categorical factors should have linear relationships with the log relative hazard.

Table 5 presents the results from the MFP analysis; for all categorical factors, such as Treatment, Sex and Race, the MFP procedure did suggest that the original form; for the two continuous variables, Age and MAP, nonlinear transformations were not suggested; for BMI, the following nonlinear transformation was suggested, $\mathrm{BMI}^3 + \mathrm{BMI}^3 \log \mathrm{BMI}$.

Table 5. MFP Suggested Transformations using Cox Regression Model

Factors	Age	MAP	BMI	Treatment	Sex	Race
Suggested Transformation	Age	MAP	BMI^3+BMI^3log(BMI)	Treatment	Sex	Race

After the proper functional forms were determined, the next step was to investigate interaction terms and check the proportionality assumption; the procedures were similar between the Cox models with *RCS* or *FP* transformations.

For multivariate Cox regression model with *RCS* transformation, to identify the correct functional form, Martingale residuals from a fully saturated Cox PH model was plotted against each of the continuous factors. Remember, the simulated survival data was consisted of 2000 subjects, 201 subjects from the placebo group switched to active treatment sometime after 1-year post randomization, each of whom should have 2 observations. Of the original 2000 subjects, 1500 (75%) subjects were partitioned into the training set; of the 1500 subjects, 150 placebo-randomized subjects switched to active treatment and the rest of the 1350 subjects received the same treatment during the entire study, therefore a total of 1650 observations from the 1500 subjects were included in the training set. As such, the Martingale residual could be calculated on observation level or on subject level. The obtained martingale residuals were plotted against each of the continuous factors stratified by the categorical factor or categorized continuous factors; these plots were used to determine the actual nonlinear functional forms and existence of interactions for the initial survival model. In addition, Wald tests were also performed to test the existence of interactions between every pair of factors, including both continuous and categorical factors.

To determine the actual function form of the factors for the initial survival model, a tentative model with the original factors was fit to the training set; incorrect functional forms could lead to under-detection of interaction terms. Nevertheless, it was still not worth implementing a different model, with the risk of losing extra degree of freedoms. The intension was to screen potential interaction terms for initial survival model for further selections, not to determine the exact interactions for the final model. Additionally, considering intra-subject variations, the subject-level Martingale residuals were also plotted against each continuous factor, though no cross-overs were observed among the curves from different stratifications, different slopes of the lowess fit between stratifications would suggest existence of interaction.

To ensure a comprehensive initial survival mode, it is preferable to be more conservative to prescreen the factors, covariate and interactions, just to capture all

possible information (features) from the data, since a comprehensive initial model should be more likely (than a simplified model) to capture the correct relationship among factors though model selections. Thus, all possible (including the confirmed or questionable) pairwise interactions should be included in the initial survival model for further evaluations. For any factors included in interactions, the entire functional forms of the selected factors or the categorical factors should be used to construct the interaction terms.

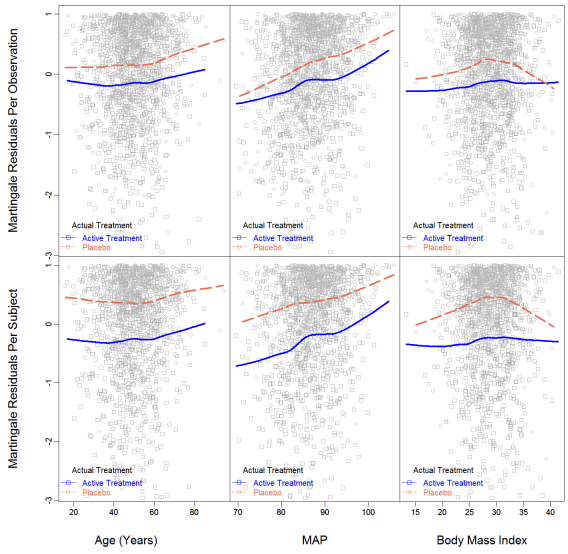


Figure 11. Martingale Residuals against Continuous Factors Stratified by Treatment – Simulation Study

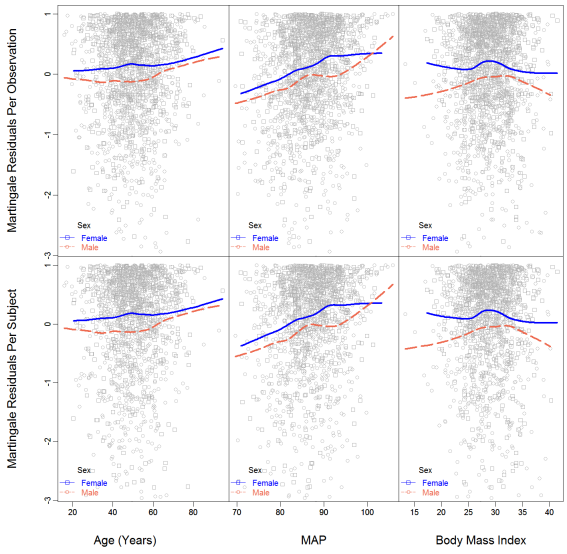


Figure 12. Martingale Residuals against Continuous Factors Stratified by Sex – Simulation Study

Figure 11 displays the Martingale residuals vs. each of the continuous factors stratified by the actual Treatment groups. As mentioned earlier, residuals could be obtained from each observation; the observation-level residuals are presented in the top 3 plots of the figure; 3 inflection points were observed for MAP; for Age and BMI, one inflection point was observed from the curves of lowess fit. After adjustment of multiple observations per subject, the subject-level residuals were obtained; results are presented in the bottom 3 plots. In the subject-level residual plot, 3 inflection points were observed for Age and MAP, 1 inflection point was observed for BMI; to be more conservative, 5-knot *RCS* transformation for Age and MAP, 3-knot *RCS* transformation for BMI was considered in the initial model with *RCS* transformations.

However, for the Cox model with *FP* transformation, the actual functional form for

each factor was already determined in Table 5 and only the suggested functional form was considered for the initial model with *FP* transformations. In addition, from the residual plot, the smoothed lowess curves of Age were parallel between the treatment groups, which suggested no interactions between Age and treatment. As for MAP, the lowess plots for different treatment groups were not crossed, but different slopes were apparent between the two treatment groups, which was an evidence for the existence of interactions, although the evidence was not substantial. Further investigation was attempted with hypothesis test of the interaction term involving Treatment and MAP. As for BMI, the placebo group had shown slight quadratic pattern in the Martingale residuals, but pretty much straight for the group with Active treatment, however the overall pattern did not show different slope, which suggested no interactions.

Table 6. Wald Test on Interactions between Treatment and All Other Factors – Simulation Study

	χ^2	df	P-value
Treatment	74.82	6	<.0001
All Interactions	9.07	5	0.1065
Age	3.11	2	0.2109
With Treatment	0.03	1	0.8553
MAP	54.28	2	<.0001
With Treatment	5.54	1	0.0186
BMI	1.23	2	0.5408
With Treatment	0.41	1	0.5204
Sex	31.64	2	<.0001
With Treatment	1.32	1	0.2511
Race	25.56	2	<.0001
With Treatment	2.51	1	0.1130
Total Interaction	9.07	5	0.1065
TOTAL	175.73	11	<.0001

Table 7. Wald Test on Interactions between Sex and All Other Factors – Simulation Study

	χ^2	df	P-value
Sex	38.32	6	<.0001
All Interactions	7.74	5	0.1715
Age	3.56	2	0.1688
With Sex	0.17	1	0.6784
MAP	49.37	2	<.0001
With Sex	0.86	1	0.3536
BMI	1.78	2	0.4103
With Sex	1.00	1	0.3182
Treatment	68.47	2	<.0001
With Sex	1.37	1	0.2412
Race	27.95	2	<.0001
With Sex	4.85	1	0.0276
Total Interaction	7.74	5	0.1715
TOTAL	170.54	11	<.0001

To further investigate the interactions involving Treatment factor, a tentative model with all interactions involving treatment was fit to the data. Results are presented in Table 6, where a significant *P*-value for the interaction term should be a strong evidence for confirmation of interaction, but a non-significant *P*-value did not confirm non-interaction, it just means that there was not enough evidence to confirm interaction. Combining the evidence from the hypothesis testing and the Martingale residual plots, the decision

should be much more reliable; it was unlikely to make wrong decisions with both evidences. Age, BMI and Sex had no interactions with Treatment, but MAP and Race might have interactions with Treatment.

Similarly, the interactions between Sex and all other variables were also examined; the plots of Martingale residual are presented in Figure 12 and the interactions terms were tested using Wald test; results are presented in Table 7. From Figure 12, it can be seen that the lowess fitted curves for different sexes were parallel for Age and BMI, but were crossed for MAP, which suggested interactions between MAP and Sex; controversially, such interaction was not significant from the Wald tests (*P*-value=0.3536). Instead, the Wald tests found significant interaction between Sex and Race. Looking more carefully at the figure, it can be seen that curve of MAP for male and female pointed in different directions (concave and convex), which suggested nonlinear pattern of the factor, and such nonlinear pattern could lead to the under-detection of the interactions from the Wald test. Considering these clues, the questionable interactions between Sex and MAP, and between Sex and Race should both be considered for the initial model, just to be conservative.

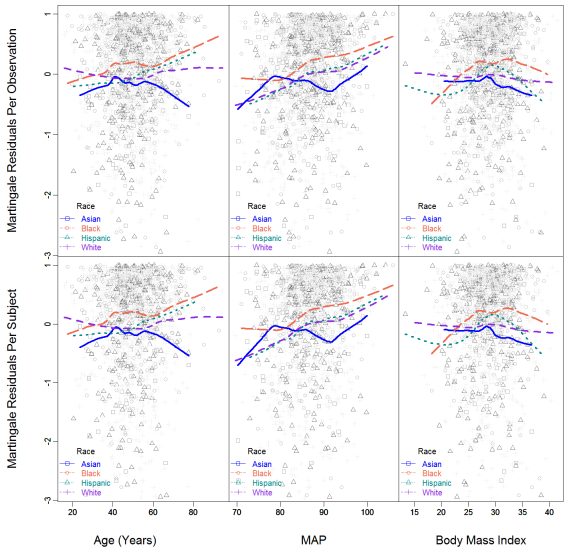


Figure 13. Martingale Residuals against Continuous Factors Stratified by Race – Simulation Study

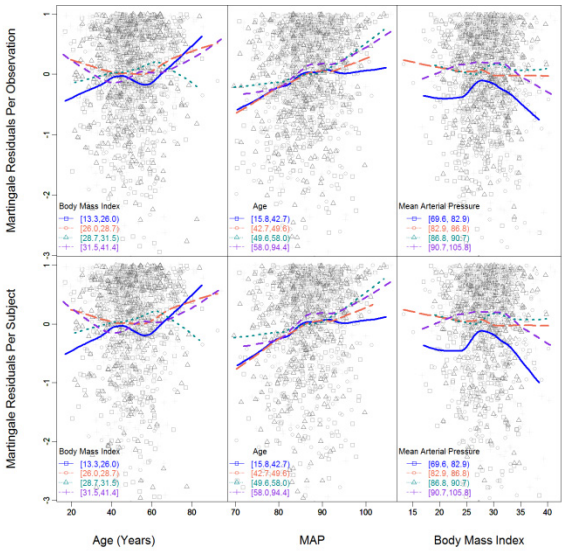


Figure 14. Martingale Residuals for Continuous Factors Stratified by the Quantiles of Another – Simulation Study

The interactions between Race and all other factors were also investigated; residual plots are presented in Figure 13; the lowess fitted curves for different Races were pretty

much parallel most of the time, but there were a few crossovers between different Races in all of the three continuous factors, and many inflection points were also observed. Therefore, it was suspected that the few crossovers could be due in part to the nonlinear factors or due in other part to the random chance from small number of subjects; no substantial evidence had suggested interactions of any of the continuous factors with Race; the other factors considered for potential interactions with Race was Sex and BMI. Specifically, the lowess curves for Asian subjects were crossed with those from other Races, but looking at the small number of Asian subjects from Table 1, the evidence did not stand out substantially.

Further evaluations were performed through hypothesis testing via Wald test; results are presented in Table 8. The results of Wald tests only suggested the interaction between Race and Sex (*P*-value = 0.0263). For interaction between Race and BMI, Wald test was boundary significant (*P*-value =0.0556), thus Race might be interacted with BMI, which should be considered for the initial model, if the design matrix was invertible. Just to be more conservative, the interaction between Treatment and Race (*P*-value=0.1118) were also considered if the model was estimable.

Table 8. Wald Test on Interactions between Race and All Other Factors – Simulation Study

	χ^2	d.f.	P-value
Race	37.68	6	<.0001
All Interactions	12.84	5	0.0249
Age	5.24	2	0.0728
With Race	1.62	1	0.2028
MAP	48.65	2	<.0001
With Race	0.18	1	0.6726
BMI	4.34	2	0.1144
With Race	3.66	1	0.0556
Treatment	70.23	2	<.0001
With Race	2.53	1	0.1118
Sex	36.69	2	<.0001
With Race	4.94	1	0.0263
Total Interaction	12.84	5	0.0249

Table 9. Wald Test on Pairwise Interactions between Continuous Factors – Simulation Study

	χ^2	df	P-value
Age	6.06	3	0.1088
All Interactions	1.22	2	0.5423
MAP	43.95	2	<.0001
With Age	1.02	1	0.3124
BMI	0.39	2	0.8234
With Age	0.07	1	0.7909
Total Interaction	1.22	2	0.5423
MAP	44.00	3	<.0001
All Interactions	1.24	2	0.5381
BMI	0.41	2	0.8161
With MAP	0.09	1	0.7660
Total Interaction	1.24	2	0.5381

Tests on interactions involving at least one categorical factor were relatively easy to perform; but the interactions between continuous factors were not very straightforward with the residual plots, since no stratification could be utilized. Thus, one factor had to be

categorized into intervals using the cutoff points of 25, 50 and 75 percentiles. The residuals plots for each continuous factor stratified by the categorized factor are presented in Figure 14. The lowess residual curves were parallel for most of the time among different intervals, but there were a few inflections points and a few crossovers, which could be due to actual interactions or due to small number of subjects within each interval. Further confirmation could be achieved based on the *P*-values from the Wald test (hypothesis testing on interactions), which are presented in Table 9. There was no evidence for considering interactions between the continuous factors.

4.1.2.2.4.2 Proportionality

As shown previously in 4.1.2.2.4 that the non-proportionality was properly addressed after adjusting for the time-varying treatment effect with extension option 1 (see section 3.8.5 for details) for the linear model. However, just in case the non-proportionality was not hidden due to the improper functional forms of the covariates, the proportionality assumptions were further assessed again for all covariates using the functional forms as identified from section 4.1.2.2.4.1. At this step, Cox regression model with all possible functional forms as well as the potential interactions as identified from the previous section was tentatively fit to the survival data. The tentative Cox model with RCS transformations is shown below.

$$\begin{aligned}\lambda(t|X) = {} & \beta_1^T RCS(\text{Age}, 5) + \beta_2^T RCS(\text{MAP}, 5) + \beta_3^T RCS(\text{BMI}, 3) + \\ & \beta_4^T \text{I}(\text{Treatment}) + \beta_5^T \text{I}(\text{Sex}) + \beta_6^T \text{I}(\text{Race}) + \\ & \beta_7^T RCS(\text{MAP}, 5) : \text{I}(\text{Treatment}) + \\ & \beta_8^T RCS(\text{MAP}, 5) : \text{I}(\text{Sex}) + \beta_9^T RCS(\text{BMI}, 3) : \text{I}(\text{Race}) + \\ & \beta_{10}^T \text{I}(\text{Treatment}) : \text{I}(\text{Race}) + \beta_{11}^T \text{I}(\text{Sex}) : \text{I}(\text{Race})\end{aligned} \quad \text{.....Eq. 56}$$

where RCS with u-knots was already defined in Eq. 48 from section 3.8.3, note that the u-knot *RCS* transformations should have $u + 1$ terms, therefore for each *RCS* transformation, the corresponding regression coefficient, β, was a vector with $u + 1$ entries (but only $u - 1$ df were spent). The symbol I, is the information function, where $\text{I}(Z) \equiv I_{Z_i} = \begin{cases} 1, \text{if } Z = \text{Level } i \\ 0, \text{Otherwise} \end{cases}$, $i = 1, \dots, h - 1$ and h is the total number of categories for the nominal factor Z (categorical factor with 3 or more category levels).

Race had 4 categories: 3 dummy variables were created corresponding to the relative difference compared to the reference level, therefore the coefficient, β_6, should have 3

entries corresponding to the 3 dummy variables. For Sex and Treatment, only 2 categories were observed for either variables, no dummy variables were needed, and the coefficient, β_4 and β_5, should be scalar corresponding to the relative difference between the two categories; yet in the formula, the regression coefficients were presented using vector forms, with only 1 entry.

Another tentative model with respect to fractional polynomial should include the same factors as the model with *RCS* transformation, except that the covariate terms included the nonlinear form of continuous factors as identified from Table 5, the original categorical factors as well as all possible interactions as previously discussed. The tentative model is shown below.

$$\begin{aligned}\lambda(\mathrm{t}|\mathrm{X}) = {} & \alpha_1^{\mathrm{T}} \times \mathrm{Age} + \alpha_2^{\mathrm{T}} \times \mathrm{MAP} + \alpha_3^{\mathrm{T}} \times [\mathrm{BMI}^3 + \mathrm{BMI}^3 \log(\mathrm{BMI})] + \alpha_4^{\mathrm{T}} \times \\ & \mathrm{Treatment} + \alpha_5^{\mathrm{T}} \times \mathrm{Sex} + \beta_6^{\mathrm{T}} \times \mathrm{Race} + \alpha_7^{\mathrm{T}} \times \mathrm{MAP{:}\,Treatment} + \\ & \alpha_8^{\mathrm{T}} \times \mathrm{MAP{:}\,Sex} + \alpha_9^{\mathrm{T}} \times [\mathrm{BMI}^3 + \mathrm{BMI}^3 \log(\mathrm{BMI})]{:}\,\mathrm{Race} + \alpha_{10}^{\mathrm{T}} \times \\ & \mathrm{Treatment{:}\,Race} + \alpha_{11}^{\mathrm{T}} \times \mathrm{Sex{:}\,Race}\end{aligned} \quad \text{..... Eq. 57}$$

where coefficients are represented using α_i, where $i = 1 \ldots 11$, just to be differentiated from the Cox model with *RCS* transformations.

Proportionality assumption for each predictor was examined using the Cox model with *RCS* transformation as presented in Eq. 56; the Scaled Schoenfeld residual plots are presented in Figure 15 and the results of the Wald tests for proportionality assumptions are presented in Table 10. Figure 15 displays the first 200 scaled Schoenfeld residuals along with the fitted least square lines (in orange), a reference line at Beta(t)=0, smoothed spline curves (in black); the number of 200 was chosen so that the smoothed spline curves were representative of the survival data, but the individual residuals should not hide the pattern of the spline curves. None of the predictors showed substantial evidence for non-proportionality, which was confirmed by the χ^2 tests presented in Table 10. The table presents the Pearson product-moment correlation between the scaled Schoenfeld residuals and time for each covariate, the χ^2 statistics for the proportionality test for each predictor, and the corresponding *P*-values. Based on the results from the table, there was not enough evidence to suggest violation of the proportionality assumption. Therefore, time-dependency was not considered for the Cox model with *RCS* transformations after the time-varying treatment effect was adjusted with the AG extension.

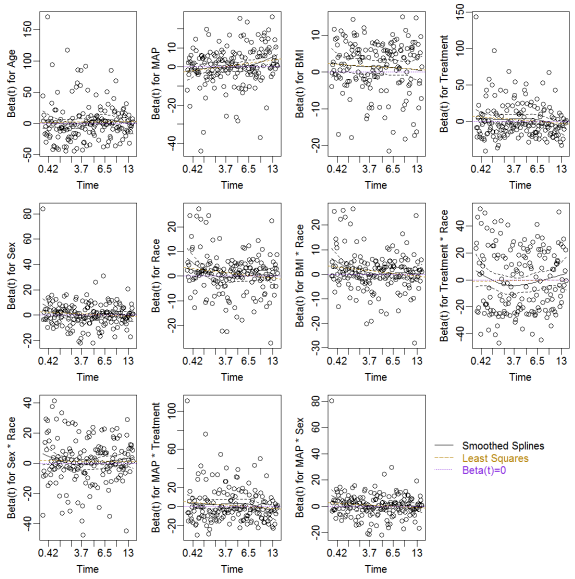


Figure 15. Scaled Schoenfeld Residual Plot for Each Covariate Based on the Cox Model with RCS Transformation – Simulation Study

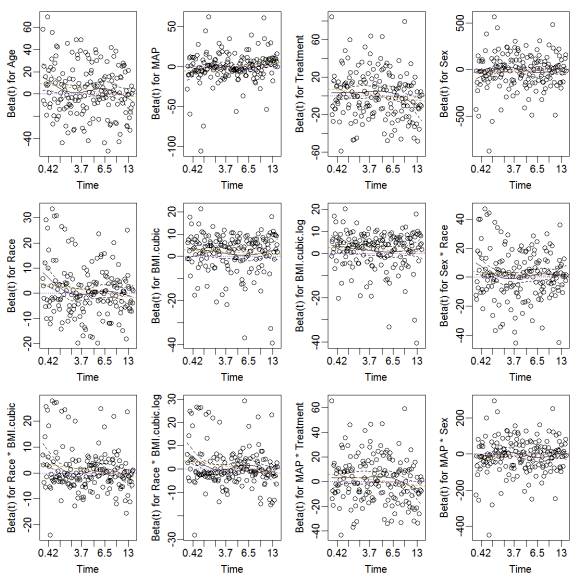


Figure 16. Scaled Schoenfeld Residual Plot for Each Covariate Based on the Cox PH Model with FP Transformation – Simulation Study

For *FP* transformation, a tentative Cox PH model was attempted including factors with suggested fractional polynomial transformations (see Table 5 for details) as well potential interactions as previously identified. Figure 16 shows the plot of the Scaled Schoenfeld residuals for each covariate for the Cox model with *FP* transformation.

Table 10. Proportionality Test on Each of the Predictor in the Cox PH model with *RCS* Transformation – Simulation Study

Factors	ρ	χ^2	*P*
Age	-0.0112	0.16	0.687
MAP	-0.0012	0.00	0.963
BMI	0.0010	0.00	0.971
Treatment	-0.0363	1.85	0.174
Sex	-0.0190	0.57	0.448
Race	-0.0191	0.50	0.480
BMI:Race	-0.0206	0.58	0.446
Treatment:Race	0.0032	0.01	0.906
Sex:Race	0.0094	0.12	0.726
MAP:Treatment	-0.0328	1.51	0.220
MAP:Sex	-0.0194	0.60	0.438
GLOBAL	NA	9.35	0.589

Table 11 presents the correlation coefficient of the scaled Schoenfeld residual with time for each covariate; a significant nonzero correlation coefficient is a strong evidence

for non-proportionality. It can be seen from Table 11, Treatment showed boundary significance against proportionality assumption (p-value =0.088); in Figure 16, the lowess curve of the scaled Schoenfeld residual for Treatment was declining at the tail, which indicated that the time-dependency for treatment effect was not completely resolved. Other than Treatment, none of the other predictors showed any evidence against proportionality assumption. The cause for this non-proportionality was possibly due to inclusion of the nonlinear functional forms of BMI; although the non-proportionality was only boundary significant, still it had to be addressed just to be conservative.

For the above reason, a time-dependent interaction terms should be further attempted following extension 2 for the Cox model with *FP* transformations after adjustment of time-varying treatment effect (extension 1 as described in section 3.8.5); the interaction term was constructed between Treatment and the corresponding treatment duration, which was further added to the covariates for the Cox model as formulated in Eq. 57. The Treatment Duration was transformed as:

$$\text{Duration.TF} = \begin{cases} \text{Duration}, & \text{if Duration} \leq 2 \text{ Yrs} \\ \exp(-\text{Duration}^2), & \text{if Duration} > 2 \text{ Yrs} \end{cases}.$$

Table 11. Proportionality Test on Each of the Covariate for the Cox model with *FP* Transformation – Simulation Study

	$\boldsymbol{\rho}$	$\boldsymbol{\chi^2}$	$\boldsymbol{P}$
Age	-0.0114	0.19	0.666
MAP	0.0273	1.04	0.308
Treatment	-0.0448	2.92	0.088
Sex	0.0406	2.34	0.126
Race	-0.0238	0.77	0.379
BMI^3	0.0081	0.10	0.758
BMI^3log(BMI)	0.0077	0.08	0.771
Sex:Race	-0.0004	0.00	0.988
BMI^3:Race	-0.0255	0.90	0.344
BMI^3log(BMI):Race	-0.0247	0.84	0.360
MAP:Treatment	-0.0416	2.51	0.113
MAP:Sex	0.0413	2.42	0.119
GLOBAL	NA	14.00	0.302

Proportionality assumptions were tested again for the Cox model with *FP* transformation including the interaction between Treatment and Duration.TF (extension option 2) with adjustment of the time varying Treatment effect (extension option 1); the

scaled Schoenfeld residual plots for all covariates are presented in Figure 17.

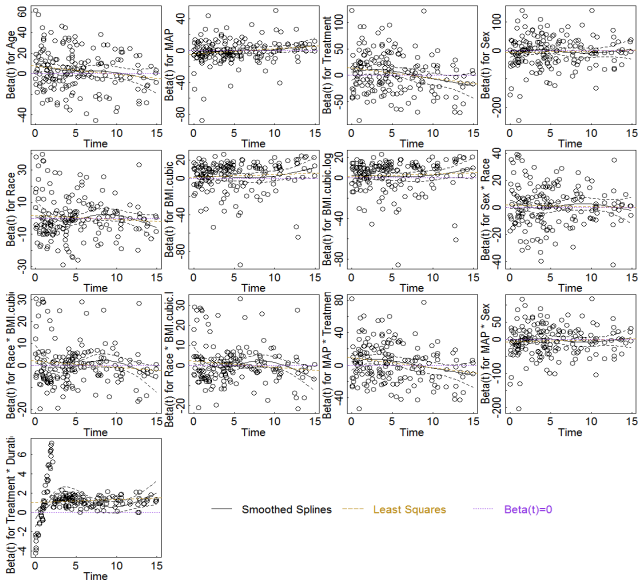


Figure 17. Scaled Schoenfeld Residual Plot for Each Covariate for the Cox Model with *FP* Transformation including the Interaction between Treatment and Transformed Treatment Duration – Simulation Study

Table 12. Proportionality Test on Each of the Predictor in the Cox PH model with *FP* Transformation (Including Interaction between Treatment and Transformed Treatment Duration) – Simulation Study

	ρ	χ^2	P
Age	-0.0287	1.19	0.274
MAP	0.0366	1.89	0.169
Treatment	-0.0378	2.05	0.152
Sex	0.0380	2.03	0.154
Race	0.0007	0.00	0.979
BMI^3	0.0134	0.26	0.608
BMI^3log(BMI)	0.0146	0.31	0.577
Sex:Race	-0.0028	0.01	0.917
BMI^3 :Race	-0.0096	0.13	0.718
BMI^3log(BMI):Race	-0.0128	0.23	0.632
MAP:Treatment	-0.0348	1.74	0.187
MAP:Sex	0.0394	2.19	0.139
Treatment:Duration.TF*	-0.0545	2.47	0.116
GLOBAL	NA	13.90	0.378

$$* \text{ Duration. TF} = \begin{cases} \text{Duration}, & \text{if Duration} \leq 2 \text{ Yrs} \\ \exp(-\text{Duration}^2), & \text{if Duration} > 2 \text{ Yrs} \end{cases}$$

The proportionality tests of all covariates, including the time-dependent interaction between treatment and treatment duration for the Cox model with *FP* transformation are

presented Table 12. It can be seen from the table that the non-proportionality was resolved (p-value=0.152), but comparing with Figure 17, the pattern of the residual for the time-dependent treatment effect (the interaction between treatment and transformed treatment duration) was still prominent within the first 2 years. However, looking at the entire study period, the severity of non-proportionality had been alleviated to a reasonable level so that the Cox model with *FP* transformation could be implemented with adjustment of the time varying treatment effect as well as the time-dependent effect in terms of the interaction between the treatment and the transformed duration.

4.1.2.3 Analysis (Model Selection)

So far data from the simulation study was ready for analysis; factors were properly reduced and clustered, model assumptions were carefully checked, appropriate functional forms and interactions between factors were identified. Next, different survival models could be attempted.

4.1.2.3.1 Conventional Cox Regression

The Cox linear model was carried out including the linear form of all factors in their original scale and the 2nd order interactions as identified in section 4.1.2.2.4.1; this model was considered as a reference for comparison with all other models. The multiple observations due to Treatment switching were adjusted using Andersen-Gill (AG) extension (extension 1 from Section 3.8.5 adjusting the time-varying treatment effect). The formula for the initial Cox PH model is shown below; all interactions identified in section 4.1.2.2.4 were included in the model.

$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}$, where $X\hat{\beta} =$

$$\begin{aligned}&\text{Age} + \text{BMI} + \text{MAP} + \text{Treatment} + \text{Sex} + \text{Race} + \text{Treatment: Race}\\&+ \text{MAP: Treatment} + \text{Sex: Race} + \text{MAP: Sex} + \text{BMI: Race}\end{aligned}$$

Models were selected using backward step-down procedure with AIC as the selection rule. Intentionally, AIC ≥1e-10 was preset so that all factors could be deleted eventually and AIC for all models were traced. Table 13 presents the summary of all deleted factors in the order of deletion from the Cox PH linear model; the first column showed the names of the covariates deleted; the order of the covariate names in the column was the same as it was deleted from the model; for the full model with all factors in linear forms, the total

df was 19, the sum of the dfs from all deleted factors.

Table 13. Summary of Linear Factors Deleted from Cox Linear Model with Backward Selection Using AIC as the Selection Rule – Simulation Study

Deleted	χ^2	df	P	Residual	AIC
Race	1.12	3	0.7712	1.12	-4.88
Treatment:Race	3.86	6	0.2767	4.99	-7.01
Sex	0.12	7	0.7268	5.11	-8.89
MAP:Sex	1.21	8	0.2719	6.32	-9.68
Treatment	3.04	9	0.0813	9.35	-8.65
Age	3.69	10	0.0547	13.05	6.95
BMI	4.29	11	0.0383	17.34	4.66
BMI:Race	28.14	14	<.0001	45.47	17.47
Sex:Race	29.08	17	<.0001	74.55	40.55
MAP	58.23	18	<.0001	132.78	96.78
MAP:Treatment	56.25	19	<.0001	189.03	151.03

Figure 18 displays the Model AIC vs. the degree of freedom; each deleted factor was labelled on the X-axis.

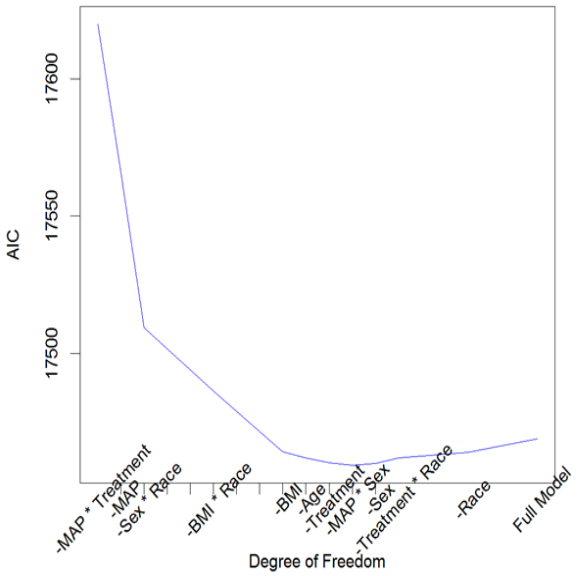


Figure 18. Model AIC vs. df of Cox Linear Model After Backward Selection – Simulation Study

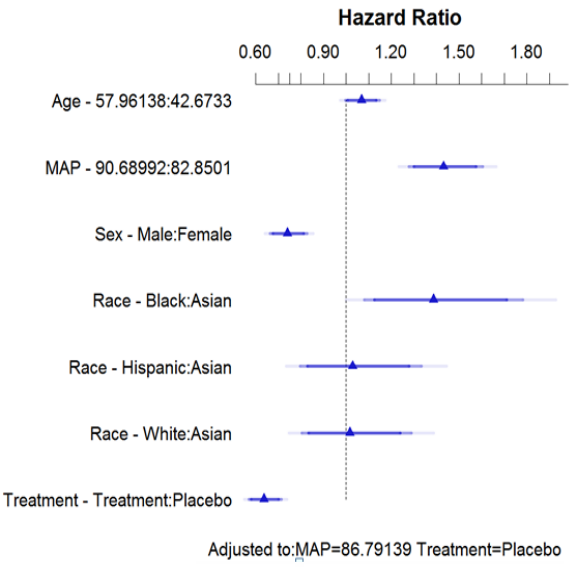


Figure 19. Hazard Ratio Obtained from the Selected Cox Linear Model – Simulation Study

Comparing the table with the figure, a good choice was to retain MAP:Treatment, MAP, Sex:Race BMI:Race, Treatment, Sex, Race, BMI and Age. As a general rule, if an interaction was retained in the model, both factors involved in the interaction should be

retained in the model; as an example, including Treatment did make slight deterioration to the model (increase in AIC), still it should be included in the final model since it interacted with MAP and the interaction between MAP and Treatment was retained in the model. According to the deletion rule mentioned in section 3.11 (the minimum number of deletions to achieve the maximum improvement in performance), both MAP and Treatment should be retained in the model; for the same reason, Sex, Race and BMI should also be included in the model. Additionally, the factor, Age, was also included in the model due to the boundary p-value. As such, only covariates (Treatment:Race, MAP:Sex) should be removed. After further looking at the final model, the interaction terms BMI:Race and Sex:Race were removed due to high p-values and the covariate term BMI was also removed for the same reason.

Table 14. Regression Coefficients of the Selected Cox Linear Model from Backward Step-Down Selection – Simulation Study

	Coef	SE	Z	Pr(>|Z|)
Age	0.004	0.0023	1.86	0.0628
{Male}	-0.301	0.0557	-5.41	<.0001
Race {Black}	0.328	0.1275	2.57	0.0102
{Hispanic}	0.029	0.1312	0.22	0.8241
{White}	0.017	0.1202	0.14	0.8894
MAP	0.046	0.0074	6.18	<.0001
{Active Treatment}	1.429	0.8407	1.70	0.0892
MAP:{Active Treatment}	-0.022	0.0096	-2.25	0.0242

At last, the regression coefficients from the Cox linear model with the selected factors, the standard errors of the coefficients and the corresponding *P*-values were estimated from the reduced model; results are presented in Table 14. It has to be cautious when interpreting the factors involved in the interaction terms, because the effect of individual factors could not be properly interpreted without accounting for the interactions. For example, Active Treatment had a regression coefficient of 1.429, but since it was involved in the interaction with MAP, the coefficient of Treatment was only referred to as the log hazard of Active Treatment vs. Placebo given MAP=0. Therefore for better interpretation, the log hazard and hazard ratios were estimated with each continuous factors fixed at the median and categorical factors fixed at the lowest alphabetic level; results are presented in Table 15. Please note that the estimated hazard ratio of Male vs. Female was -0.298 (Table 15), which was slightly different from the

coefficients (-0.301) for Male relative to female as presented in Table 14; this was possible due to rounding errors. Additionally, for continuous factors, the inter-quartile hazard ratios and the corresponding 95% CI are 9 presented in Table 15 and the forest plots of the estimated hazard ratios corresponding to the regression coefficients obtained from the selected Cox linear model are displayed in Figure 19.

For continuous factor, the low and high columns corresponds to the lowest and highest values for the factor; the difference is the difference between the highest and lowest values for the factor; the effect, is the log hazard/hazard ratios corresponding to unit change in the factor; SE is the standard error of the log hazard corresponding to the unit change in the factor; lower and upper 95% CI are the corresponding lower and upper bounds of the 95% CIs. The hazard ratio was obtained by taking the exponential of the regression coefficients, so were the lower and upper bound of the 95% CIs. For categorical variables, the effect is the log relative hazard (or hazard ratios) between the category levels. In this table, the interaction effects are not presented separately from the individual factors, instead they are summarized by different levels of the factors involved in the interactions: using Treatment as an example, interaction existed between MAP and Treatment (see Table 14); the hazard ratio between Active Treatment and Placebo was calculated based on the median MAP.

Table 15. Regression Coefficients and the Corresponding Hazard Ratio Estimates from the Selected Cox Linear Model after Backward Selection – Simulation Study

	Low	High	Diff.	Effect	SE of Effect	Lower 95% CI	Upper 95% CI
Age	42.67	57.96	15.29	0.067	0.0357	-0.0034	0.1365
Hazard Ratio	42.67	57.96	15.29	1.069		0.9966	1.1463
MAP	82.85	90.69	7.84	0.388	0.0743	0.2427	0.5341
Hazard Ratio	82.85	90.69	7.84	1.475		1.2747	1.7060
Sex - Male:Female	1	2		-0.298	0.0559	-0.4081	-0.1889
Hazard Ratio	1	2		0.742		0.6649	0.8278
Race - Black:Asian	1	2		0.330	0.1276	0.0801	0.5801
Hazard Ratio	1	2		1.391		1.0834	1.7863
Race - Hispanic:Asian	1	3		0.035	0.1315	-0.2228	0.2926
Hazard Ratio	1	3		1.036		0.8003	1.3400
Race - White:Asian	1	4		0.020	0.1204	-0.2157	0.2561
Hazard Ratio	1	4		1.020		0.8060	1.2919
Treatment - Active:Placebo	1	2		-0.452	0.0572	-0.5645	-0.3403
Hazard Ratio	1	2		0.636		0.5686	0.7116

After the model was selected, the model was cross validated using 10-fold cross validation (CV) as described in section 3.9; the performance statistics obtained from the 10-fold leave-one-out cross validation are presented in Table 16. Index.Orig was referring to the performance statistics estimated from the entire training set; training was referring to the statistics estimated from the 9-fold of the bootstrapped dataset within the CV step, and test was referring to the 1-fold sample that were left out from each iteration of CV; optimism was the difference in the performance statistics between the training and testing within the CV step. Index.corrected was the value obtained by subtracting the optimism from Index.Orig. A positive optimism indicates overfit and a negative optimism indicates underfit. The statistics presented in the table do not suggest underfit or overfit; the selected model was the best fit for the data for Cox PH linear model.

Table 16. CV Performance of the Selected Cox Linear Model from the Backward Selection – Simulation Study

	Index.Orig	Training	Test	Optimism	Index.Corrected
Dxy	-0.1979	-0.1987	-0.1874	-0.0113	-0.1866
R2	0.0927	0.0933	0.0878	0.0055	0.0872
Slope	1.0000	1.0000	0.9671	0.0329	0.9671
D	0.0091	0.0093	0.0125	-0.0032	0.0123
U	-0.0001	-0.0001	0.0010	-0.0011	0.0010
Q	0.0092	0.0094	0.0115	-0.0021	0.0113
g	0.4266	0.4282	0.4113	0.0169	0.4097

After the Cox linear model was trained and cross validated, it was fit to the test set for evaluation the prediction performance, in terms of time-dependent prediction errors and time-dependent AUCs; results are summarized in Table 17. The prediction errors were calculated using the sum of the squared difference between the predicted survival probability and the actual survival status of each subject at the corresponding time point; smaller prediction error suggested better predictions. The AUC was the area under the ROC curve; a value of 0.5 corresponded to the probability of a random guess and a value of 1 corresponded to 100% accuracy (100% sensitivity and specificity); a value of 65% is meaningful, value of 70% suggests a good fit and a value of 75% indicates excellent prediction. The plots of the prediction errors and time-dependent AUCs as well as the corresponding 95% PCIs are presented and Figure 20.

Table 17. Prediction Performance of the Cox Linear Model – Simulation Study Test Set

Yrs	Prediction Error (95% PCI)	AUC (95% PCI)
1	0.1118 (0.0933, 0.1329)	0.6832 (0.6194, 0.7463)
2	0.1674 (0.1494, 0.1849)	0.6438 (0.5968, 0.6923)
3	0.2021 (0.1872, 0.2163)	0.6316 (0.5896, 0.6749)
4	0.2177 (0.2065, 0.2276)	0.6307 (0.5957, 0.6707)
5	0.2324 (0.2224, 0.2421)	0.6218 (0.5868, 0.6587)
6	0.2350 (0.2236, 0.2471)	0.6207 (0.5889, 0.6543)
7	0.2330 (0.2190, 0.2476)	0.6190 (0.5887, 0.6530)
8	0.2290 (0.2120, 0.2466)	0.6176 (0.5883, 0.6489)
9	0.2214 (0.2013, 0.2433)	0.6171 (0.5878, 0.6484)
10	0.2096 (0.1835, 0.2325)	0.6173 (0.5877, 0.6477)
11	0.2006 (0.1723, 0.2278)	0.6172 (0.5881, 0.6485)
12	0.1921 (0.1610, 0.2217)	0.6178 (0.5888, 0.6482)
13	0.1877 (0.1521, 0.2223)	0.6167 (0.5878, 0.6471)
14	0.1793 (0.1405, 0.2169)	0.6163 (0.5877, 0.6467)

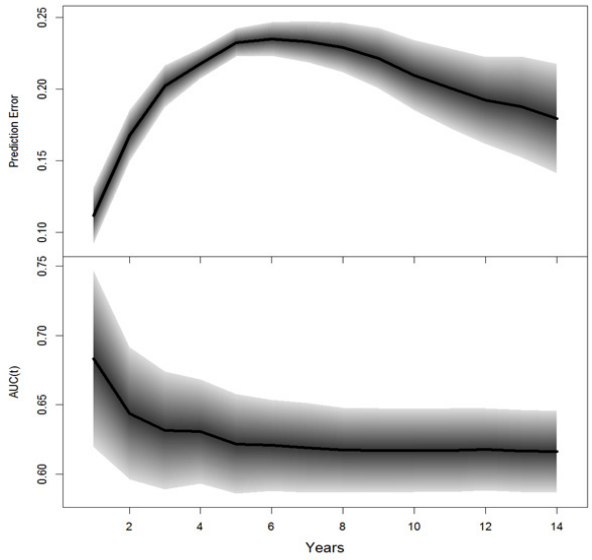


Figure 20. Prediction Performance of the Selected Cox Linear Model – Simulation Study Test Set

With the selected model, the survival probability for future events and median survival time were predicted. The nomogram, a simple way to display the predictions for an average subject, was shown in Figure 21. For any given subject, there was a point corresponding to each of the factors, then the projection of the sum of the Points from all factors to the line of total points, can be used to obtain the predicted log hazard ($X\hat{\beta}$); the projection of the sum of the Points of all factors to the lines of the 1-year, 3-year, 5-year

and 10-year survival probability or the median survival time to estimate the corresponding 1, 3, 5 and 10 year survival probability or median survival time.

At last, the log-hazard (the regression coefficients) and the survival probability vs. each continuous factor stratified by Sex and Treatment, the predicted survival probability stratified by each categorical factor, and the median survival time for each continuous factor and the corresponding 95% CI are displayed in Figure 22. Additional to the log hazard (regression coefficients), the relative hazards (or the hazard ratios) for each predictor could be derived easily from the log hazard; it was just the exponential of the log-hazard; which are not presented in this paper, just to save some space.

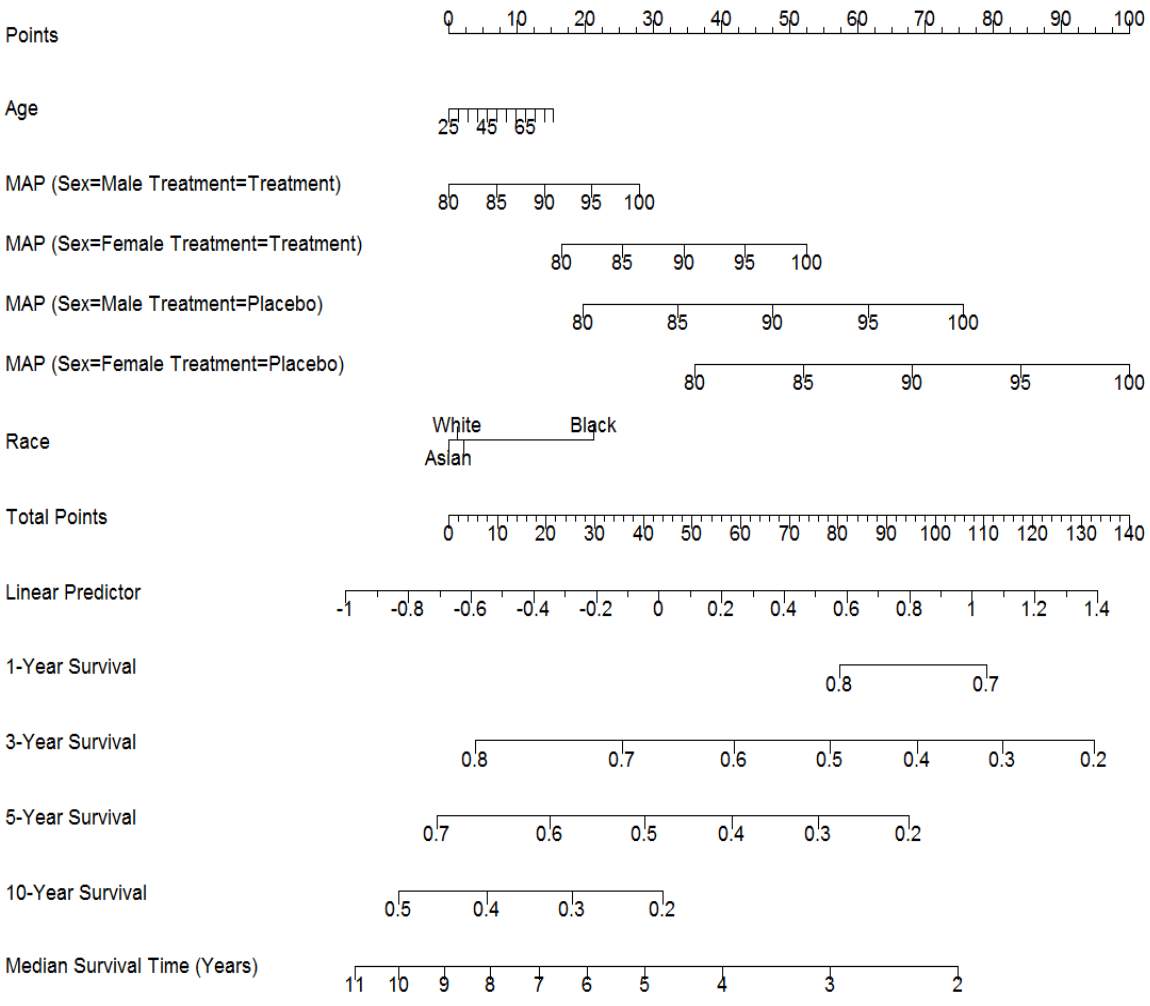


Figure 21. Nomogram of Survival Probability and Median Survival Time Based on the Selected Cox PH linear model – Simulation Study

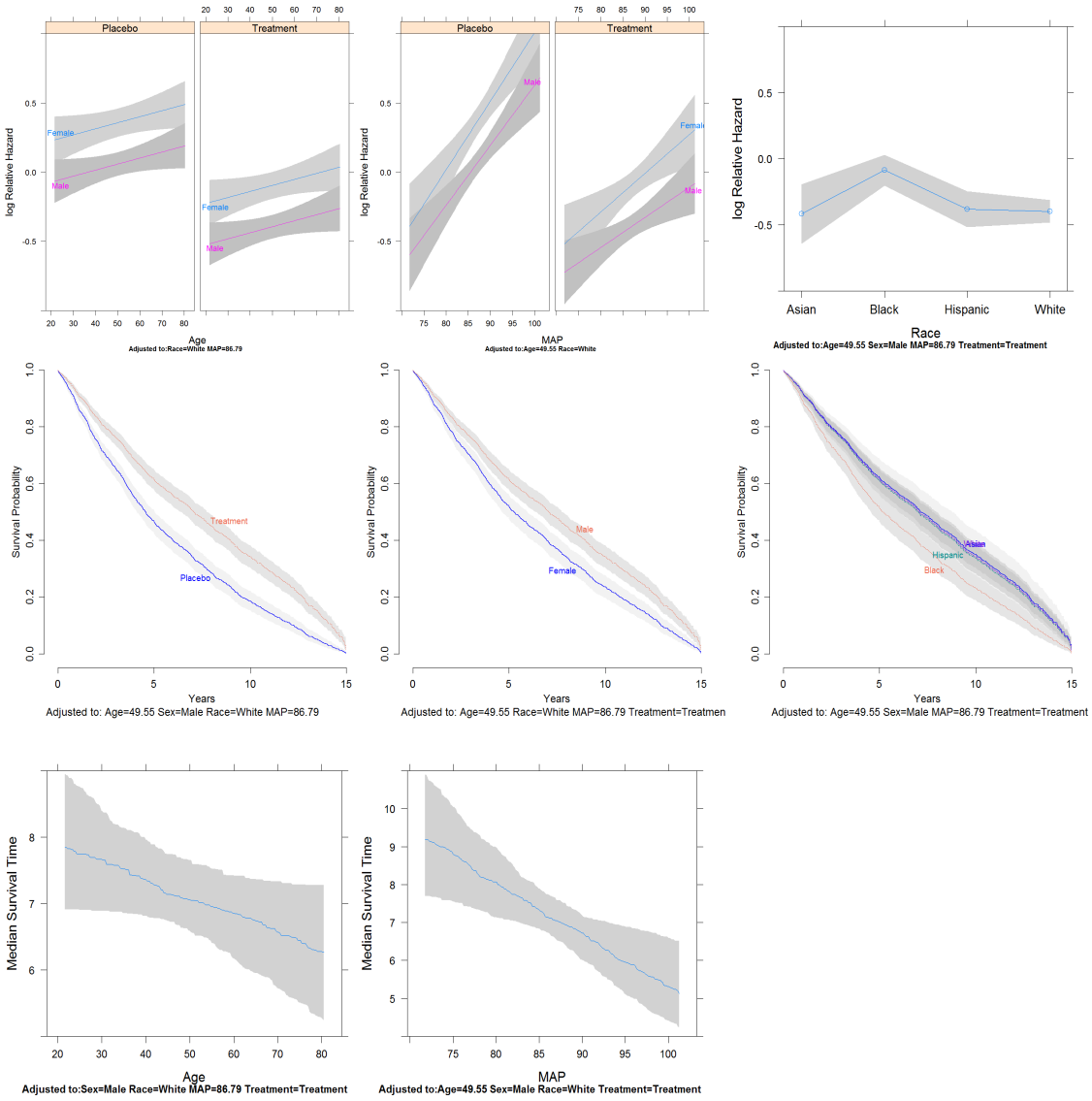


Figure 22. Plots of Log Hazard, Survival Probability and Median Survival Times based on the Selected Cox Linear Model – Simulation Study

4.1.2.3.2 Multivariate Cox Regression Models

4.1.2.3.2.1 Multivariate Cox Regression with *RCS* Transformation

The initial Cox model with *RCS* transformation was built with the *RCS* transformation of all continuous factors, the original categorical factors and all potential interaction terms as identified in section 4.1.2.2.4.1; multiple observations per subject with time-varying treatment effect were adjusted using AG counting process extension (extension option 1 from section 3.8.5). The initial model is presented below.

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \text{where } X\hat{\beta} =$$

$$\begin{aligned}
&-5.4693 + 1.5100\{\text{Treatment}\} + 4.8264\{\text{Male}\} - 0.4968\{\text{Black}\} - 0.7431\{\text{Hispanic}\} + 0.8408\{\text{White}\} \\
&-0.0072\text{Age} + 2.5691\times10^{-5}(\text{Age} - 29.9398)_+^3 - 0.0002(\text{Age} - 43.2900)_+^3 + 0.0002(\text{Age} - 49.5460)_+^3 \\
&\quad -0.0001(\text{Age} - 56.9116)_+^3 + 2.2273\times10^{-5}(\text{Age} - 70.6781)_+^3 \\
&+0.0630\text{MAP} - 0.0003(\text{MAP} - 76.7173)_+^3 + 0.0023(\text{MAP} - 83.3120)_+^3 - 0.0042(\text{MAP} - 86.7914)_+^3 \\
&\quad +0.0025(\text{MAP} - 90.1678)_+^3 - 0.0004(\text{MAP} - 96.5421)_+^3 \\
&+0.0293\text{BMI} - 0.0008(\text{BMI} - 23.7275)_+^3 + 0.0016(\text{BMI} - 28.6586)_+^3 - 0.0008(\text{BMI} - 33.6296)_+^3 \\
&+\{\text{Male}\}[-0.3060\ \{\text{Black}\} + 0.0899\ \{\text{Hispanic}\} - 0.0269\ \{\text{White}\}] \\
&+\{\text{Treatment}\}[-0.0233\text{MAP} + 0.0002(\text{MAP} - 76.7173)_+^3 - 0.0011(\text{MAP} - 83.3120)_+^3 + 0.0018(\text{MAP} - 86.7914)_+^3 \\
&\quad -0.0009(\text{MAP} - 90.1678)_+^3 + 6.1611\times10^{-5}(\text{MAP} - 96.5421)_+^3] \\
&+\{\text{Male}\}[-0.0651\text{MAP} + 0.0012(\text{MAP} - 76.7173)_+^3 - 0.0094(\text{MAP} - 83.3120)_+^3 + 0.0185(\text{MAP} - 86.7914)_+^3 \\
&\quad -0.0125(\text{MAP} - 90.1678)_+^3 + 0.0022(\text{MAP} - 96.5421)_+^3] \\
&+\{\text{Treatment}\}[-0.1592\ \{\text{Black}\} - 0.0766\ \{\text{Hispanic}\} + 0.0581\ \{\text{White}\}] \\
&+\{\text{Black}\}[0.0367\text{BMI} + 0.0002(\text{BMI} - 23.7275)_+^3 - 0.0005(\text{BMI} - 28.6586)_+^3 + 0.0002(\text{BMI} - 33.6296)_+^3] \\
&+\{\text{Hispanic}\}[0.0289\text{BMI} - 6.1736\times10^{-5}(\text{BMI} - 23.7275)_+^3 + 0.0001(\text{BMI} - 28.6586)_+^3 - 6.1239\times10^{-5}(\text{BMI} - 33 \\
&+\{\text{White}\}[-0.0340\text{BMI} + 0.0006(\text{BMI} - 23.7275)_+^3 - 0.0013(\text{BMI} - 28.6586)_+^3 + 0.0006(\text{BMI} - 33.6296)_+^3]
\end{aligned}$$

Similar to what was done for Cox PH linear model, the above model was backward selected following a step-down procedure using AIC as the selection rule; again AIC>=1e-10 was preset to ensure all terms to be deleted. The summary of all deleted covariates from the multivariate Cox Regression model with *RCS* transformations are presented in Table 18.

Table 18. Summary of Backward Selection for Cox Model with *RCS* Transformed Factors – Simulation Study

Deleted	χ^2	df	P	Residual	AIC
BMI:Race	8.61	6	0.1969	8.61	-3.39
Treatment:Race	2.33	3	0.5060	10.94	-7.06
MAP:Treatment	5.26	4	0.2616	16.2	-9.8
Sex:Race	5.33	3	0.1493	21.53	-10.47
Age	2.94	2	0.2294	24.47	-11.53
Sex	2.71	1	0.0995	27.19	10.81
BMI	6.77	2	0.0339	33.96	8.04
MAP	22.75	4	0.0001	56.71	6.71
Race	24.67	3	0.0000	81.38	25.38
Treatment	69.41	1	0.0000	150.79	92.79
MAP:Sex	79.43	4	0.0000	230.22	164.22

The plot of AIC of the remaining model vs. the degree of freedom after each deletion for Cox model with *RCS* Transformations are displayed in Figure 23; each deleted factor was labelled on the X-axis.

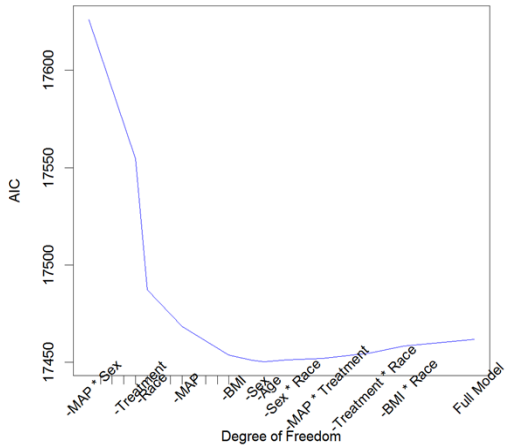


Figure 23. Model AIC vs. df after Each Backward Deletion for Cox Model with *RCS* Transformations – Simulation Study

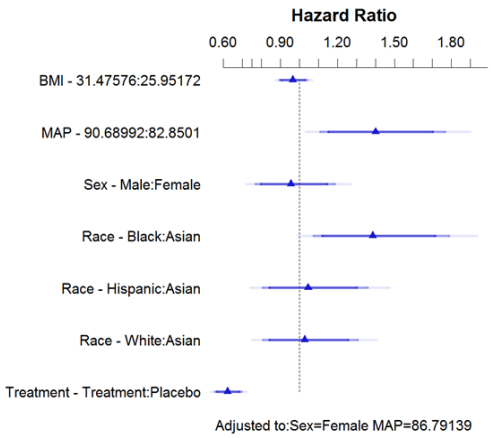


Figure 24. Hazard Ratios Estimated from the Selected Cox Model with RCS Transformations – Simulation Study

The reduced Cox model with *RCS* transformation as selected from the backward step down procedure is summarized in Table 19; the formulation of the reduced model is:

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \text{where } X\hat{\beta} =$$
$$-5.5101 + 6.24306\{\text{Male}\} + 0.3262\{\text{Black}\} + 0.0454\{\text{Hispanic}\} + 0.0282\{\text{White}\} - 0.4754\{\text{Treatment}\}$$
$$+0.0260\text{BMI} - 0.0004(\text{BMI} - 23.7275)_+^3 + 0.0009(\text{BMI} - 28.6586)_+^3 - 0.0004(\text{BMI} - 33.6296)_+^3$$
$$+0.0625\text{MAP} - 0.0004(\text{MAP} - 76.7173)_+^3 + 0.0025(\text{MAP} - 83.3120)_+^3 - 0.0046(\text{MAP} - 86.7914)_+^3$$
$$+0.002977986(\text{MAP} - 90.1678)_+^3 - 0.0005(\text{MAP} - 96.5421)_+^3 + \{\text{Male}\}[-0.0837\text{MAP} + 0.00138(\text{MAP} - 76.7173)_+^3$$
$$-0.0106(\text{MAP} - 83.3120)_+^3 + 0.0202(\text{MAP} - 86.7914)_+^3 - 0.0134(\text{MAP} - 90.1678)_+^3 + 0.0023(\text{MAP} - 96.5421)_+^3]$$

Table 19. Regression Coefficients of the Selected Cox Model with *RCS* Transformations from Backward Step-Down Selection – Simulation Study

	Coef	S.E.	Z	P(>|Z|)
Sex=Male	6.243	3.8531	1.62	0.1052
Race=Black	0.326	0.1293	2.52	0.0116
Race=Hispanic	0.045	0.1328	0.34	0.7326
Race=White	0.028	0.1214	0.23	0.8169
Treatment=Active	-0.475	0.0574	-8.28	<.0001
BMI	0.026	0.0151	1.73	0.0844
BMI'	-0.043	0.0177	-2.43	0.0149
MAP	0.063	0.0417	1.50	0.1337
MAP'	-0.140	0.1816	-0.77	0.4403
MAP''	0.990	1.0592	0.93	0.3498
MAP'''	-1.824	1.7453	-1.05	0.2960
Sex=Male : MAP	-0.084	0.0487	-1.72	0.0859
Sex=Male : MAP'	0.544	0.2209	2.46	0.0137
Sex=Male : MAP''	-4.148	1.3237	-3.13	0.0017
Sex=Male : MAP'''	7.952	2.2253	3.57	0.0004

The summary of the coefficients (/or log hazards) and the corresponding hazard ratios for all covariates estimated from the Cox model with *RCS* transformations after backward step-down selection is presented in Table 20. The forest plot of the hazard ratio corresponding to each of the factor is displayed in Figure 24; the hazard ratio for each factor was estimated using the above model with all other factors fixed.

Table 20. Coefficients and Hazard Ratios of All Covariates from the Selected Cox Model with RCS Transformations After Backward Selection – Simulation Study

	Low	High	Diff.	Effect	SE (Effect)	Lower 0.95	Upper 0.95
BMI	25.95	31.48	5.52	-0.0359	0.0381	-0.1106	0.0387
Hazard Ratio	25.95	31.48	5.52	0.9647		0.8953	1.0395
MAP	82.85	90.69	7.84	0.3373	0.1183	0.1054	0.5692
Hazard Ratio	82.85	90.69	7.84	1.4012		1.1112	1.7669
Sex - Male:Female	1	2		-0.0460	0.1103	-0.2621	0.1701
Hazard Ratio	1	2		0.9550		0.7694	1.1854
Race - Black:Asian	1	2		0.3262	0.1293	0.0728	0.5797
Hazard Ratio	1	2		1.3857		1.0755	1.7854
Race - Hispanic:Asian	1	3		0.0454	0.1328	-0.2148	0.3056
Hazard Ratio	1	3		1.0464		0.8067	1.3574
Race - White:Asian	1	4		0.0281	0.1214	-0.2098	0.2661
Hazard Ratio	1	4		1.0285		0.8107	1.3048
Active:Placebo	1	2		-0.4754	0.0574	-0.5879	-0.3629
Hazard Ratio	1	2		0.6216		0.5555	0.6956

The performance statistics obtained from 10-fold CV are presented in Table 21. Again, there was no indication of overfit or underfit for this model.

Table 21. Model Performance of the Selected Cox Model with *RCS* Transformation – Simulation Study

	index.orig	training	test	optimism	index.corrected
Dxy	-0.2101	-0.2111	-0.1951	-0.0160	-0.1941
R2	0.1063	0.1073	0.0968	0.0105	0.0959
Slope	1.0000	1.0000	0.9373	0.0627	0.9373
D	0.0105	0.0108	0.0140	-0.0033	0.0138
U	-0.0001	-0.0001	0.0009	-0.0011	0.0009
Q	0.0106	0.0109	0.0131	-0.0022	0.0128
g	0.4593	0.4617	0.4313	0.0304	0.4288

The prediction performance of the selected Cox model with *RCS* transformations was evaluated based on the test set; the prediction errors, the time-dependent AUC and the

corresponding 95% PCIs are summarized in Table 22; the corresponding plots are displayed in Figure 25.

Table 22. Prediction Performance of the Selected Cox PH model with *RCS* Transformations – Simulation Study Test Set

Yrs	Prediction Error (95% PCI)	AUC (95% PCI)
1	0.1108 (0.0905, 0.1315)	0.6647 (0.6005, 0.7244)
2	0.1690 (0.1511, 0.1859)	0.6266 (0.5776, 0.6763)
3	0.2047 (0.1907, 0.2187)	0.6166 (0.5751, 0.6584)
4	0.2210 (0.2103, 0.2328)	0.6177 (0.5783, 0.6554)
5	0.2370 (0.2253, 0.2484)	0.6081 (0.5715, 0.6438)
6	0.2386 (0.2250, 0.2510)	0.6090 (0.5744, 0.6456)
7	0.2375 (0.2218, 0.2534)	0.6083 (0.5754, 0.6417)
8	0.2339 (0.2147, 0.2522)	0.6062 (0.5743, 0.6380)
9	0.2269 (0.2045, 0.2493)	0.6054 (0.5738, 0.6368)
10	0.2143 (0.1897, 0.2391)	0.6048 (0.5743, 0.6354)
11	0.2047 (0.1770, 0.2335)	0.6047 (0.5743, 0.6349)
12	0.1962 (0.1678, 0.2277)	0.6054 (0.5748, 0.6353)
13	0.1913 (0.1597, 0.2247)	0.6047 (0.5748, 0.6348)
14	0.1823 (0.1473, 0.2193)	0.6046 (0.5746, 0.6344)

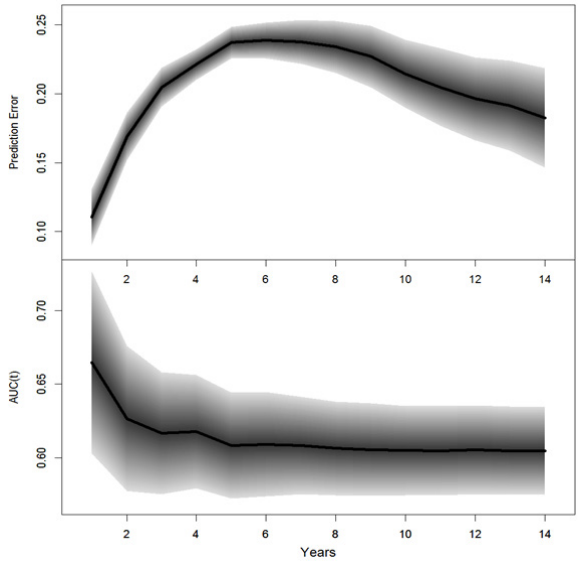


Figure 25. Prediction Errors and Time-Dependent AUCs of the Selected Cox PH Model with RCS Transformations vs. Time – Simulation Study Test Set

With the cross validated model, the log-hazards, the predicted survival probability stratified by each of the categorical factors and the predicted median survival time can be obtained, the corresponding plots are displayed in Figure 26.

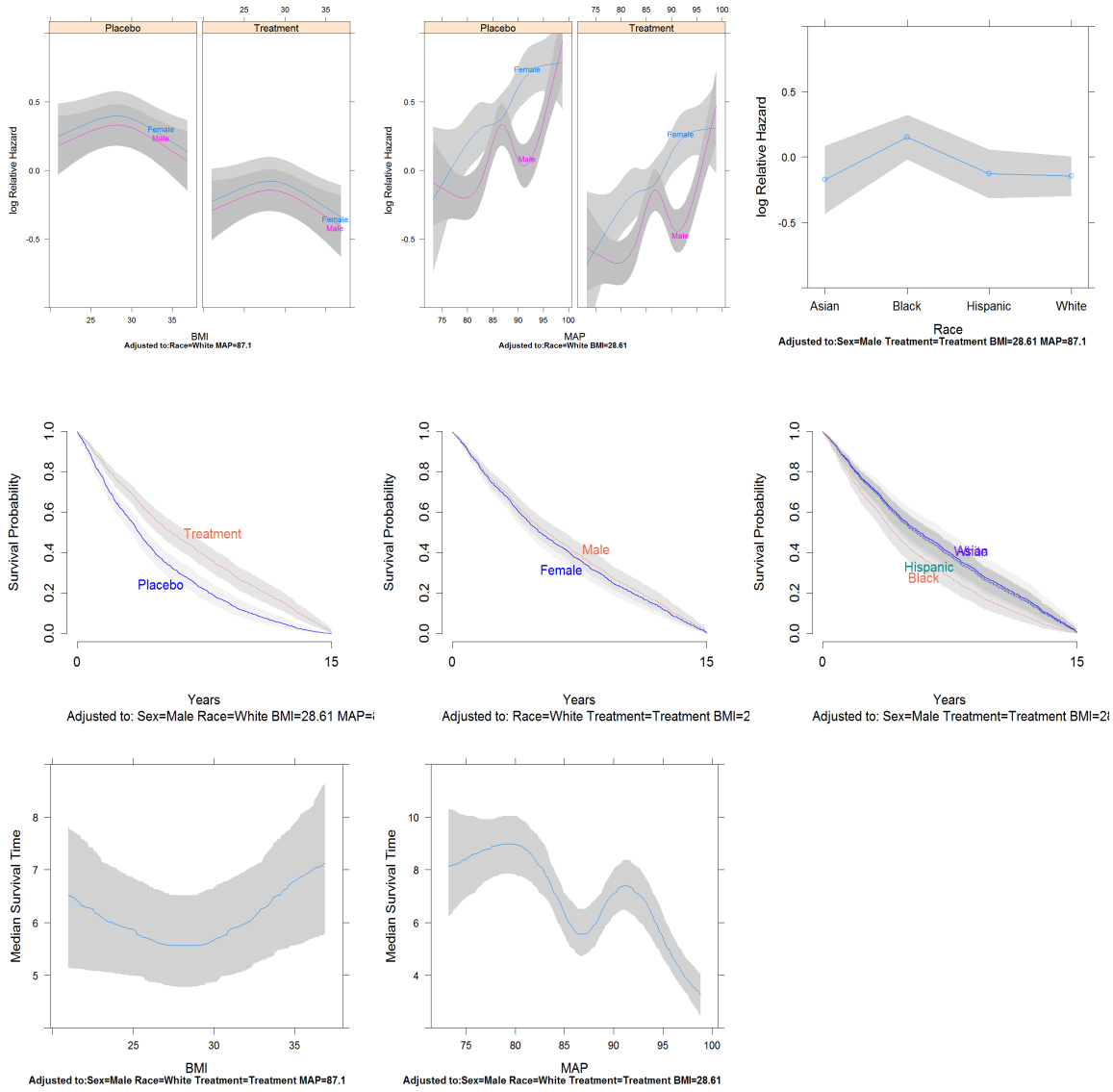


Figure 26. Predicted Log Hazard, Survival Probability and Median Survival Time based on the Selected Cox PH Model with *RCS* Transformations – Simulation Study

The nomogram of the predicted survival probability and median survival time using the selected Cox PH model with *RCS* transformation is displayed in Figure 27.

4.1.2.3.2.2 Multivariate Cox Regression Model with *FP* Transformation

The Cox model with *FP* transformation was initially attempted including the *FP* forms of continuous factors (see Table 5), interactions terms and time-dependent treatment effect as identified in section 4.1.2.2.4.1; multiple observations per subject due to time-varying treatment effect were adjusted using AG extension (extension option 1 from section 3.8.5).

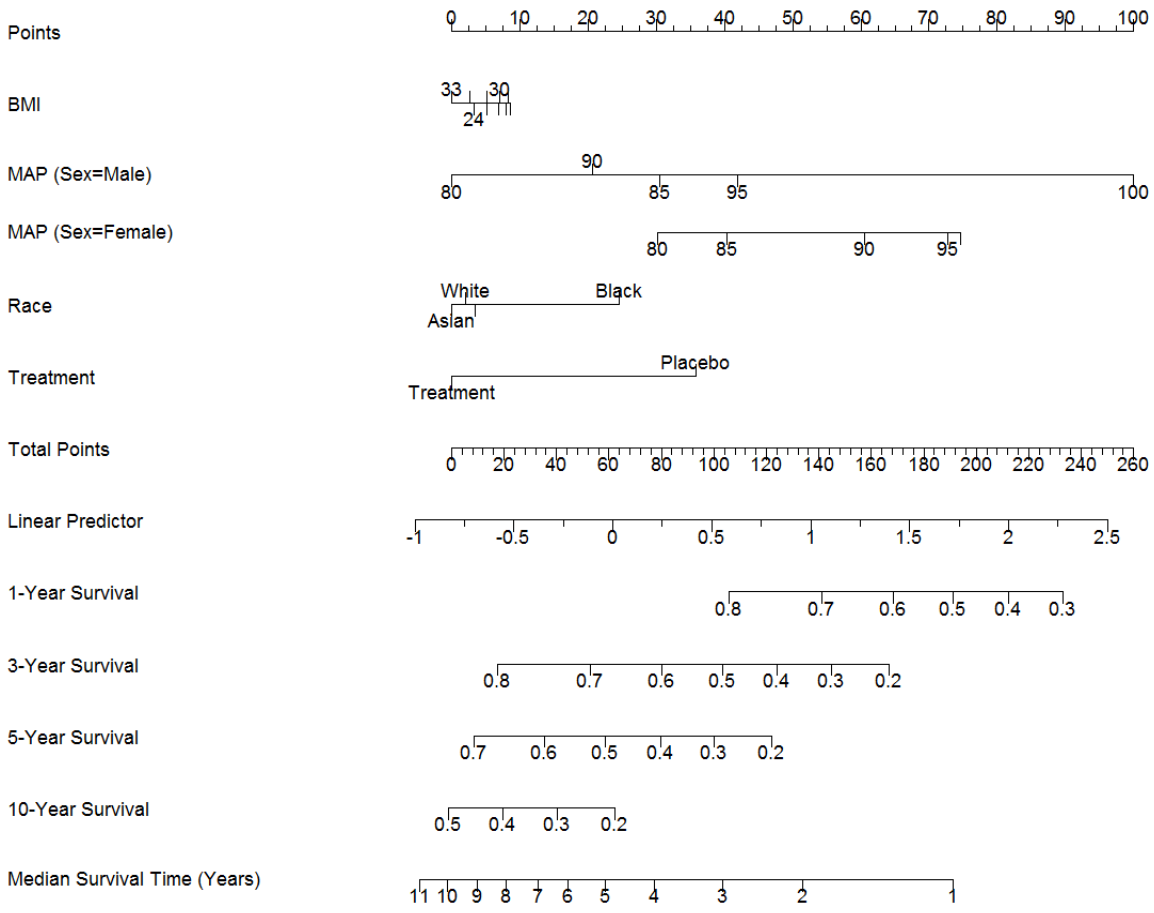


Figure 27. Nomogram of Predicted Survival Probability and Median Survival Time on Test Set with Regression coefficients Estimated from the Selected Cox Model with RCS Transformed Factors – Simulation Study

The formulation of the initial model is presented below.

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \text{where } X\hat{\beta} =$$
$$-4.2418 + 0.0049\,\text{Age} + 0.0445\,\text{MAP} - 0.9044\,\{\text{Treatment}\} + 0.6893\,\{\text{Male}\} -$$
$$0.1344\,\{\text{Black}\} - 0.5685\,\{\text{Hispanic}\} + 0.3059\,\{\text{White}\} + 0.1147\,\text{BMI}^3 -$$
$$0.0890\text{BMI}^3\log(\text{BMI}) + \text{Treatment} \times \left\{\text{ifelse}\left(\text{Duration} \leq 2, \text{Duration}, \text{e}^{-\text{Duration}^2}\right)\right\} +$$
$$\text{Male} \times [-0.3799\,\{\text{Black}\} + 0.0098\,\{\text{HIspanic}\} - 0.0770\,\{\text{White}\}] - 0.0168\,\text{MAP} \times$$
$$\{\text{Treatment}\} + \text{BMI}^3 \times [0.00561\,\{\text{Black}\} + 0.0828\,\{\text{Hispanic}\} - 0.0773\,\{\text{White}\}] +$$
$$\text{BMI}^3\log(\text{BMI}) \times [-0.0301\,\{\text{Black}\} + 0.2885\,\{\text{Hispanic}\} + 0.0897\,\{\text{White}\}] -$$
$$0.0098\,\text{MAP} \times \text{Male}$$

Again, the initial model was selected following a backward step-down procedure based on AIC; to ensure all terms to be deleted, the selection criterion was preset to AIC

≥ 1e-10. The summary of the backward selection process is presented in Table 23.

Table 23. Backward Step-Down Selection of the Cox Model with *FP* Transformations – Simulation Study

Deleted	Chi-Sq	df	P	Residual	AIC
Treatment:Race	0.83	3	0.8419	0.83	-5.17
Race	1.8	3	0.6144	2.63	-9.37
Race:BMI^3log(BMI)	2.45	3	0.4841	5.09	-12.91
Sex: Race	4.93	3	0.1772	10.01	-13.99
Sex	0.78	1	0.3779	10.79	-15.21
Treatment	1.26	1	0.2619	12.05	-15.95
BMI^3	4.14	1	0.0419	16.19	-13.81
BMI^3log(BMI)	2.01	1	0.1563	18.20	-13.8
Age	4.92	1	0.0265	23.12	-10.88
Race:BMI^3	11.88	3	0.0078	35.00	-5
MAP:Sex	26.24	1	<.0001	61.24	19.24
MAP	48.66	1	<.0001	109.90	65.9
MAP:Treatment	66.49	1	<.0001	176.39	130.39
Treatment:Duration.TF*	607.54	1	<.0001	783.93	735.93

$$\text{Duration. TF}* = \begin{cases} \text{Duration}, & \text{if Duration} \le 2 \text{ Yrs} \\ \exp(-\text{Duration}^2), & \text{if Duration} > 2 \text{ Yrs} \end{cases}$$

Figure 28 displays the plot of the model AIC vs. the remaining df after each deletion of the covariates for Cox model with *FP* transformations. Apparently, MAP, MAP:Sex, MAP:Treatment and the time-dependent treatment effect (Treatment:Duration.TF) should remain in the reduced model. Other than the above factors, the individual factors involved in the interactions as selected by the procedure such as Treatment and Sex should remain in the reduced model; the other terms such as Age, Race:[BMI^3 + BMI^3log(BMI)] and [BMI^3 + BMI^3log(BMI)] were also included in the initial Cox model due to significant p-values. The model was then fitted to the training set again; it was noticed that the term Race: [BMI^3 + BMI^3log(BMI)] was not needed due to extremely large *p*-value. Thus the final model should include Age, MAP, BMI^3, BMI^3log(BMI), Sex, Treatment, MAP:Sex, MAP:Treatment, Race and the time-dependent effect as expressed using the interaction terms of Treatment:Duration.TF (transformed Treatment Duration).

The coefficients of all terms for the selected Cox model with *FP* transformation are summarized in Table 24; only log hazards (regression coefficients from the Cox model)

were estimated instead of the hazard ratios, which could be easily derived by taking the exponential of the log hazard. Figure 29 displays the forest plot of the hazard ratio corresponding to each term in the selected Cox model with *FP* transformation.

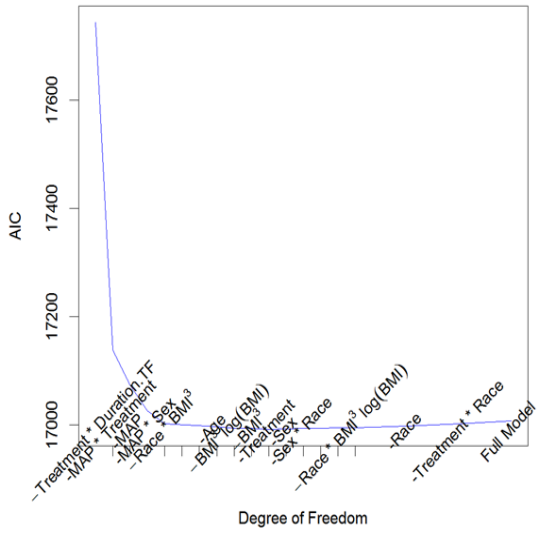


Figure 28. AIC vs. df of Backward Selection for Cox Model with FP Transformations – Simulation Study

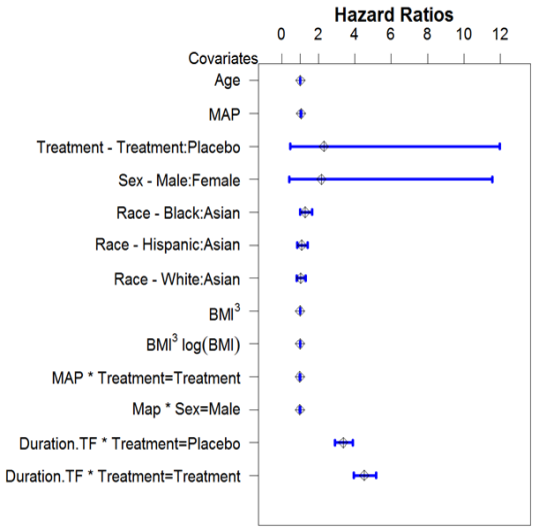


Figure 29. Hazard Ratios from the Selected Cox Model with FP Transformations – Simulation Study

Table 24. Summary of Regression Coefficients of the Selected Cox Model with *FP* Transformations – Simulation Study

	Coef	S.E.	Z	P(>|Z|)
Age	0.005	0.0023	2.05	0.0401
MAP	0.046	0.0096	4.81	0.0000
Treatment=Active	0.834	0.8406	0.99	0.3211
Sex=Male	0.775	0.8537	0.91	0.3640
Race=Black	0.252	0.1281	1.96	0.0496
Race=Hispanic	0.091	0.1317	0.69	0.4877
Race=White	0.037	0.1205	0.30	0.7614
BMI^3	0.000	0.0001	2.18	0.0296
$BMI^3 log(BMI)$	0.000	0.0000	-2.20	0.0281
MAP:{Treatment}	-0.015	0.0096	-1.59	0.1120
MAP:Sex=Male	-0.012	0.0098	-1.24	0.2144
Duration.TF*:{Placebo}	1.508	0.0704	21.42	<.0001
Duration.TF*:{Treatment}	1.215	0.0726	16.73	<.0001

$$* \ \text{Duration.TF} = \begin{cases} \text{Duration}, & \text{if Duration} \le 2 \text{ Yrs} \\ \exp(-\text{Duration}^2), & \text{if Duration} > 2 \text{ Yrs} \end{cases}$$

Previously in Figure 19 and Figure 24, the inter-quartile hazard ratios and the corresponding 95% CIs were presented for continuous covariates; but here in Figure 29, the

hazard ratio corresponding to unit increase of each continuous factors is presented. The standard errors of the hazard ratios for Treatment and Sex were much larger than the rest of the terms such that the 95% CI of the hazard ratios for the rest of the terms could hardly be seen from the graph; hence the hazard ratios for unit increase were used instead.

The formula of the selected Cox model with *FP* transformation after backward step-down selection is presented below:

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \text{where } X\hat{\beta} =$$
$$-3.5631 + 0.0048\,\text{Age} + 0.0462\,\text{MAP} - 0.88341\,\{\text{Treatment}\} + 0.7749\,\{\text{Male}\} +$$
$$0.2515\,\{\text{Black}\} + 0.0914\,\{\text{Hispanic}\} + 0.0366\,\{\text{White}\} + 0.0002\,\text{BMI}^3 -$$
$$0.0001\text{BMI}^3 \log(\text{BMI}) - 0.0153\,\text{MAP} \times \{\text{Treatment}\} - 0.0121\,\text{MAP} \times \{\text{Male}\} +$$
$$1.2146\,\text{Duration.TF} \times \{\text{Treatment}\} + 1.5083\,\text{Duration.TF} \times \{\text{Placebo}\}$$

The performance statistics for this model was evaluated using 10-fold CV; results are presented in Table 25. Again, there was no indication of overfitting or underfitting.

Table 25. Model Performance of the Selected Cox Model with *FP* Transformation – Simulation Study

	Index.Orig	Training	Test	Optimism	Index.Corrected
Dxy	-0.4100	-0.4106	-0.3976	-0.0130	-0.3970
R2	0.3196	0.3199	0.3046	0.0154	0.3042
Slope	1.0000	1.0000	0.9852	0.0148	0.9852
D	0.0360	0.0367	0.0521	-0.0154	0.0515
U	-0.0001	-0.0001	0.0004	-0.0005	0.0004
Q	0.0361	0.0368	0.0517	-0.0149	0.0511
g	0.8785	0.8798	0.8648	0.0150	0.8635

On the other hand, the prediction performance of this model was surprisingly better than the other two semi-parametric Cox regression models as discussed so far. The prediction errors and time-dependent AUCs of the selected Cox model with *FP* transformation were evaluated based on the test set at each time point; the 95% PCIs were obtained from 1000 bootstrap samples; results are summarized in Table 26. The plots of prediction errors time-dependent AUCs as well as the 95% PCIs are displayed in Figure 30.

It can be seen that the prediction performance of the selected Cox model with *FP* transformation was much better than the Cox linear models and the Cox model with *RCS* transformation. The reason for the improvement was probably due to the addition of

time-dependent interaction term of Treatment:Duration.TF, after adjustment of the time-varying treatment effect. However, this model was much more complex than the previous two models; it involved a time-dependent treatment interaction term, in which the transformation function for factor, Treatment Duration, was extremely hard to find, it had gone through a lot of trials-and-errors.

Table 26. Prediction Performance of the Selected Cox Model with *FP* Transformations – Simulation Study

Yrs	Prediction Error (95% PCI)	AUC (95% PCI)
1	0.1292 (0.1075, 0.1521)	0.7548 (0.7109, 0.7933)
2	0.1114 (0.0950, 0.1279)	0.8222 (0.7921, 0.8494)
3	0.1495 (0.1328, 0.1651)	0.7674 (0.7392, 0.7945)
4	0.1755 (0.1607, 0.1906)	0.7469 (0.7175, 0.7736)
5	0.2034 (0.1888, 0.2191)	0.7246 (0.6976, 0.7480)
6	0.2148 (0.2000, 0.2312)	0.7166 (0.6905, 0.7390)
7	0.2202 (0.2028, 0.2388)	0.7099 (0.6859, 0.7310)
8	0.2231 (0.2039, 0.2434)	0.7032 (0.6803, 0.7245)
9	0.2194 (0.1987, 0.2408)	0.6994 (0.6771, 0.7201)
10	0.2119 (0.1876, 0.2373)	0.6964 (0.6739, 0.7166)
11	0.2047 (0.1750, 0.2318)	0.6958 (0.6734, 0.7154)
12	0.1978 (0.1657, 0.2275)	0.6957 (0.6731, 0.7152)
13	0.1922 (0.1560, 0.2261)	0.6948 (0.6729, 0.7142)
14	0.1825 (0.1468, 0.2207)	0.6940 (0.6719, 0.7131)

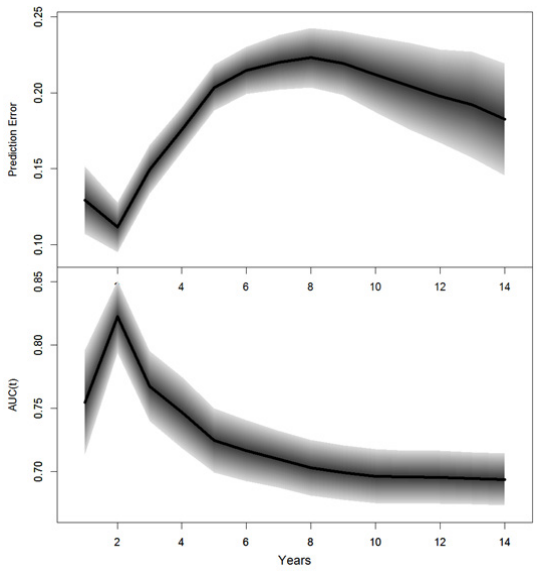


Figure 30. Time Dependent AUC vs. Time for the Selected Cox Model with *FP* Transformations – Simulation Study Test Set

With the selected Cox PH model with *FP* transformation, the log relative hazard for each factor, including Age, BMI, MAP and Race, was predicted with the rest of the factors fixed at a constant value. For interaction between continuous factor and categorical factor, separate curves were produced based on the levels of the categorical factor involved in the interaction.

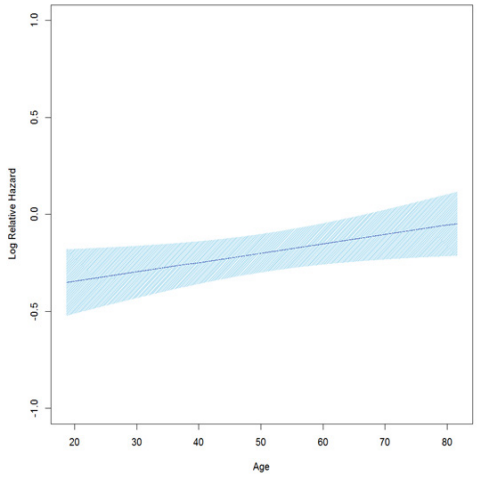


Figure 31. Predicted Log Hazard vs. Age Based on the Selected Cox Model with *FP* Transformations – Simulation Study

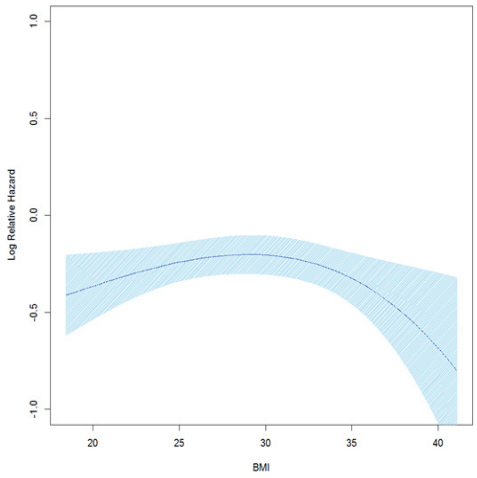


Figure 32. Predicted Log Hazard vs. BMI Based on the Selected Cox Model with *FP* Transformations – Simulation Study

Figure 31 presents the predicted log relative hazard against age and the corresponding 95% CI, with the rest of the factors fixed (continuous factors were fixed at the medians, MAP=87.1 mmHg and BMI=28.6; categorical factors were fixed at the largest category level). Changes made to any of these factors should result in a different predicted value and further lead to a different plot of log relative hazard for Age. For the same reason, a different plot could be produced for different category levels of a particular categorical factor. However, considering Age did not interact with any other factors, a separate plot was not produced, since the log relative hazard curve for Age should be parallel between levels of another factor. Similarly, BMI was a nonlinear term, but it did not interact with any other factors (see Table 24 for details), a single log relative hazard against BMI is presented in Figure 32 as well as the 95% CI (blue shaded area).

The log relative hazard for MAP was produced similarly; however the factor interacted with both Treatment and Sex; it should have different predicted log relative hazard for different treatment and different Sex, thus plots with stratification of

Treatment are presented separately for Female and Male in Figure 33. Figure 34 presents the log hazard for Race; the factor had 4 category levels, the log hazards for the 4 category levels were connected, but the actual slope of the curve was not statistically meaningful, since Race was a nominal factor; only the relative difference between levels was meaningful.

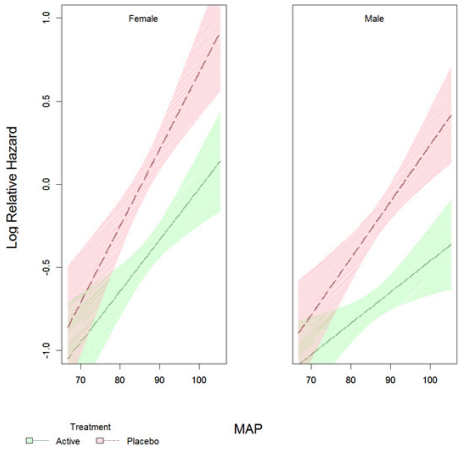


Figure 33. Predicted Log Hazard for MAP from the Selected Cox Model with *FP* Transformations – Simulation Study

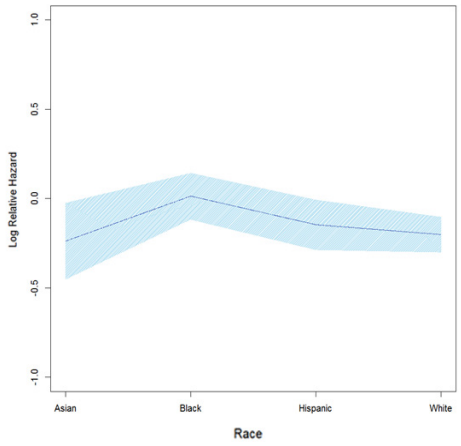


Figure 34. Predicted Log Hazard for Race from the Selected Cox Model with *FP* Transformations – Simulation Study

Figure 35 presents the predicted log relative hazard for Treatment Duration stratified by Treatment; a quadratic pattern was apparent between 0 and 3 years. Figure 36 presents the predicted median survival time against Age while fixing the rest of the factors constant.

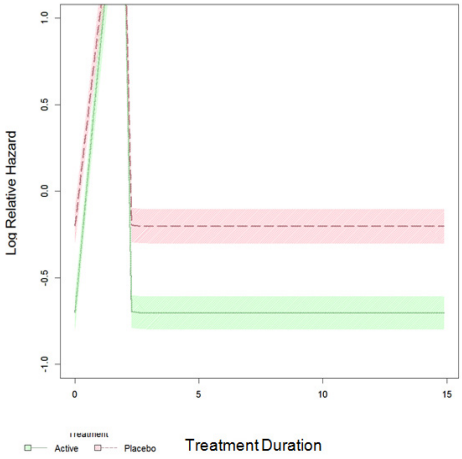


Figure 35. Predicted Log Hazard for Treatment Duration from the Selected Cox Model with *FP* Transformation – Simulation Study

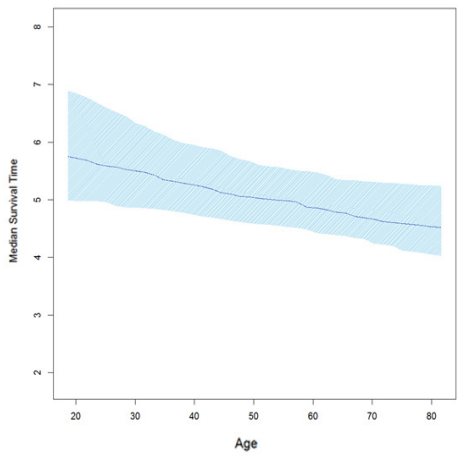


Figure 36. Predicted Median Survival Time for Age from the Selected Cox Model with *FP* Transformation – Simulation Study

Figure 37 presents the predicted median survival time vs. BMI and Figure 38 presents the median survival time vs MAP stratified by treatment for different Sex.

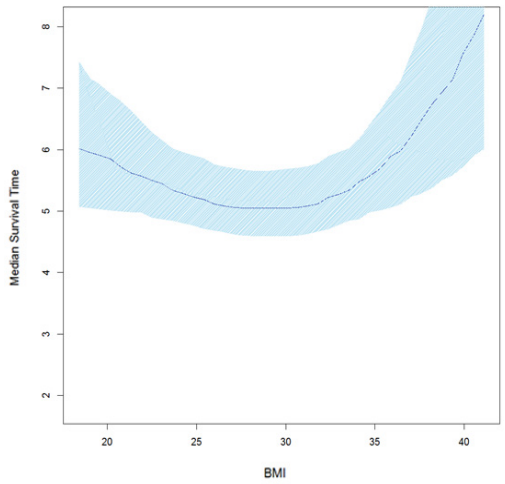


Figure 37. Predicted Median Survival Time for BMI from the Selected Cox Model with *FP* Transformation – Simulation Study

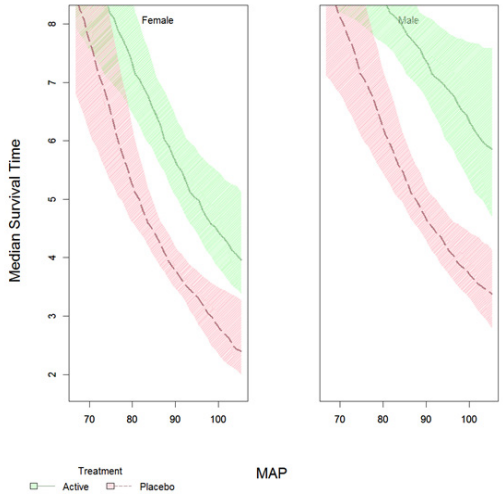


Figure 38. Predicted Median Survival Time for MAP from the Selected Cox Model with FP Transformation– Simulation Study

Figure 39 shows the median survival time vs. Race; again the slope or the incremental change had no meanings. Figure 40 displays the median survival time vs. Treatment Duration.

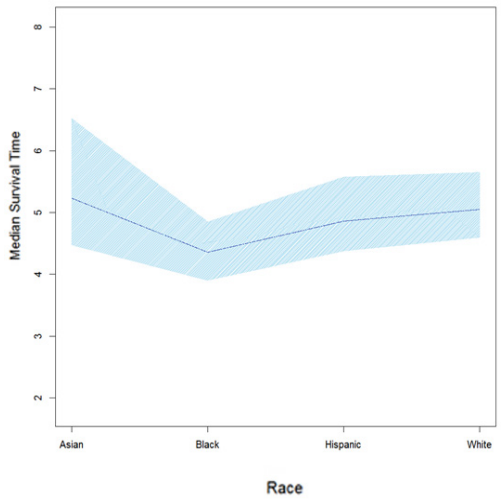


Figure 39. Predicted Median Survival Time for Race from the Selected Cox Model with *FP* Transformation – Simulation Study

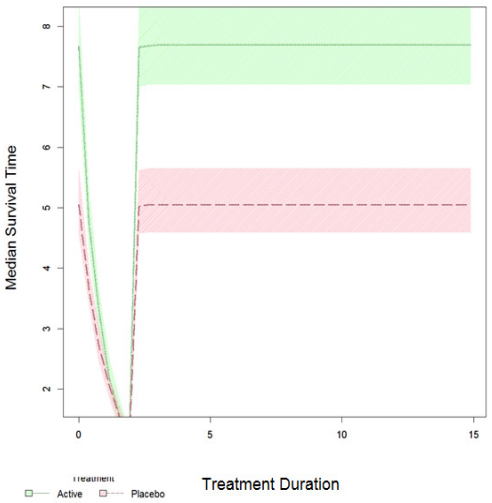


Figure 40. Predicted Median Survival Time for Treatment Duration from the Selected Cox Model with *FP* Transformation– Simulation Study

Figure 41 and Figure 42 presents the predicted survival probability stratified by Treatment for different Sex based on the selected.

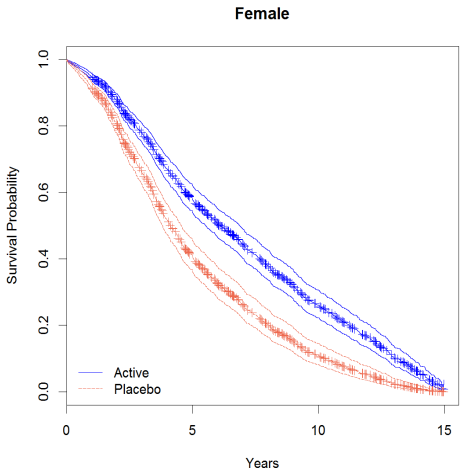


Figure 41. Predicted Survival Probability By Treatment for Female Based on the Selected Cox Model with *FP* Transformation – Simulation Study

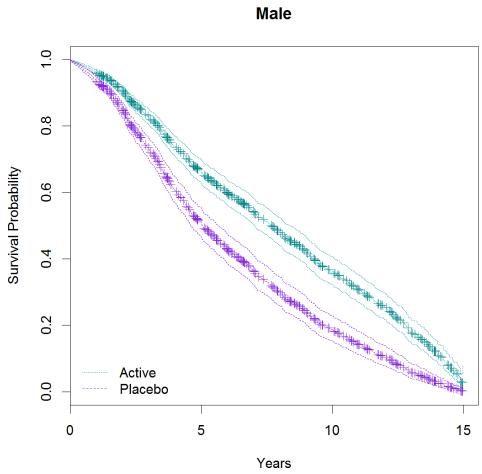


Figure 42. Predicted Survival Probability By Treatment for Male Based on the Selected Cox Model with *FP* Transformation – Simulation Study

Figure 43 and Figure 44 presents the predicted survival probability stratified by different Race for different Treatment.

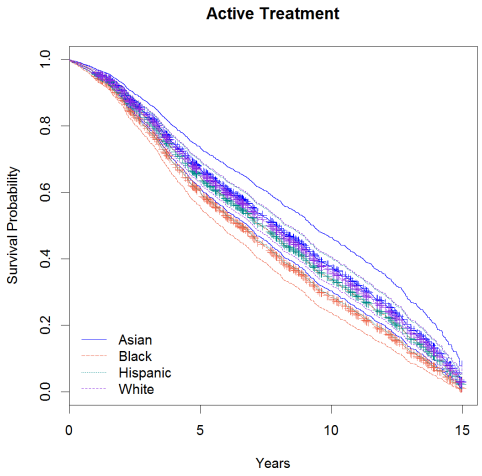


Figure 43. Predicted Survival Probability by Race for Active Treatment from the Selected Cox Model with *FP* Transformation– Simulation Study

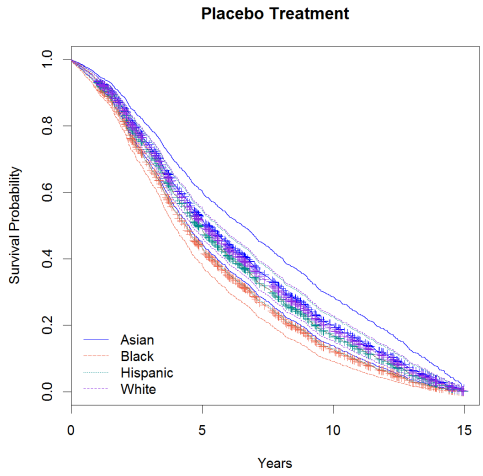


Figure 44. Predicted Survival by Race for Placebo Treated Subjects from the Selected Cox Model with *FP* Transformation – Simulation Study

4.1.2.3.3 Nonparametric Random Survival Forest (RSF)

Nonparametric, random survival forest (RSF) approaches were also introduced in

study; the approaches do not make any assumption of the survival probability or hazard function, therefore there is no need to worry about the actual functional forms for all factors, interactions or proportionality assumptions.

Two different algorithms of RSF were intended. The log-rank based RSF, had all necessary tools developed for cross validation, prediction of future outcomes, but the performance of the approach may be affected for highly correlated survival data. Additionally, the log-rank based RSF model cannot account for multiple observations per subject. Therefore in order to implement this model, the subject who switched from placebo to active treatment were considered as two different subjects, one treated with placebo starting at day 0 and censored at the time of treatment switching and the other one treated with active treatment starting at the time of treatment switching until event or censoring; this was only reasonable if the treatment switching was independent of the failure event.

A second RSF model, conditional inference (CINF) based RSF model, was also implemented for this study. The approach was developed based on conditional probability; thus it should have better performance for highly correlated data; additionally, the model is capable of handling multiple events or multiple observations per subject. However this model had just been proposed; not many features or functions were available; thus significant effort was spent to derive features and functions for evaluation of predictions and prediction performance. Additionally, a flexible function was also developed to retrieve conditional forest trees for predicting survival outcomes based on the analysis results.

4.1.2.3.3.1 Log-rank Based Random Survival Forest (RSF)

As mentioned earlier, log-rank based random survival forest model (LR-RSF) could not model multiple observations per subject due to the time-varying treatment effect; for this simulation study, placebo treated subjects who switched treatment were considered as 2 independent subjects (see section 4.1.2.1 for details), each with a different treatment for different durations; therefore this should be one of the disadvantages for the log-rank based RSF model. On the other hand, the LR-RSF is a nonparametric model and no model assumptions are involved, thus there is no need to check for nonlinearity or non-proportionality, which is one of the advantages for this approach.

The initial fit of the log-rank based RSF model was carried out including all factors; the variable importance (VIMP) and Brier scores were assessed via CV; VIMP of all factors are summarized in Table 27. Figure 45 displays the out-of-Bag (OOB) error rate and the VIMP.

Table 27. VIMP from Log-Rank Based RSF – Simulation Study

	VIMP	Relative VIMP
MAP	0.0081	1.0000
Sex	0.0080	0.9887
SBP	0.0078	0.9572
DBP	0.0056	0.6907
Race	0.0041	0.5026
BMI	0.0021	0.2646
Age	0.0004	0.0530
Treatment	0.0004	0.0501

Table 28. VIMP from Log-Rank Based RSF with SBP and DBP Removed – Simulation Study

	VIMP	Relative VIMP
MAP	0.0081	1.0000
Sex	0.0080	0.9887
Race	0.0041	0.5026
BMI	0.0021	0.2646
Age	0.0004	0.0530
Treatment	0.0004	0.0501

As previously mentioned in section 4.1.2.2.3, SBP, DBP and MAP were highly correlated, all of the three factors were equally important (as seen in Table 27 and Figure 45). Thus, a second model was attempted with SBP and DBP excluded. Table 28 presents the VIMP of all factors excluding SBP and DBP. The plots of out-of-Bag (OOB) error rate and VIMP with SBP and DBP excluded are presented in Figure 46; significant change was not observed.

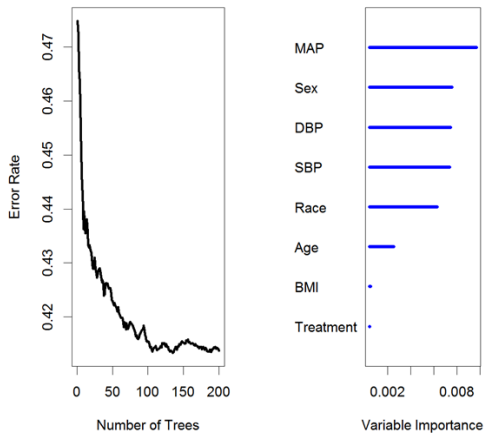


Figure 45. CV Out-of-Bag Error Rate and Variable Importance (VIMP) of Log-Rank Based RSF – Simulation Study

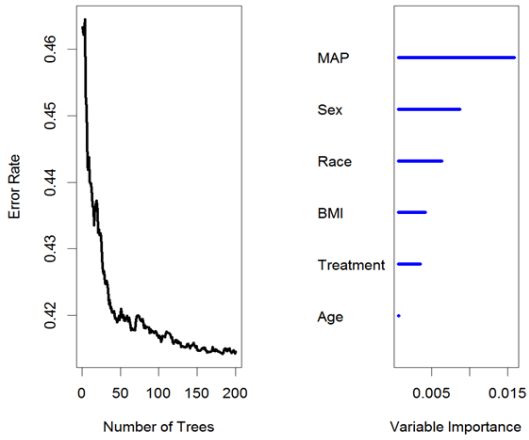


Figure 46. Out-of-Bag Error Rate and VIMP of Log-Rank Based RSF with SBP and DBP Removed – Simulation Study

However, to be consistent with all other survival models, the second model with SBP and DBP excluded was focused; with this model, pair-wise interactions were systematically checked via maximum subtree and variable importance (VIMP). Table 29 presents the normalized minimum depths for each pair of factors; for the normalized minimum depths matrix, factors with off-diagonal entries smaller than the diagonal entry, were suspected to have of interactions, thus BMI and Age were suspected to interact with all factors; Table 30 presents the analysis of all pair-wise interactions via variable importance (VIMP) in the scale of $\times 10^{-3}$; the columns of "Var1" and "Var2" are the VIMP for each pair of factors in the interaction, the column of "Paired" is the VIMP of the interaction terms, the column of "additive" is the sum of the VIMP from Var1 and Var2, and the column of "Diff" is the VIMP difference between "paired" and "additive". The interaction terms of MAP:BMI, MAP:Treatment, Sex:Treatment, Race:Treatment, and Treatment:Age seem to be important, since the absolute differences are reasonably large as compared to additive VIMP.

Table 29. Pairwise Interactions via Maximum Subtree Analysis for Log-Rank Based RSF – Simulation Study

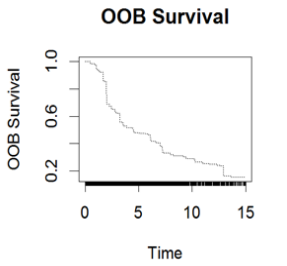
	Treat	MAP	Race	Sex	BMI	Age
Treat	0.04	0.07	0.1	0.29	0.11	0.11
MAP	0.57	0.05	0.1	0.32	0.09	0.09
Race	0.63	0.07	0.09	0.33	0.09	0.08
Sex	0.62	0.08	0.12	0.09	0.1	0.1
BMI	0.68	0.08	0.11	0.38	0.12	0.08
Age	0.68	0.08	0.11	0.4	0.08	0.13

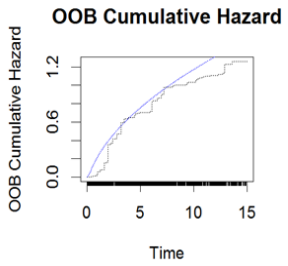
NOTE:
1. If the values in the diagonal at [i, i] entry is small and the off-diagonal elements are even smaller than the diagonal entry, then interaction may be suspected.
2. Treat: Treatment

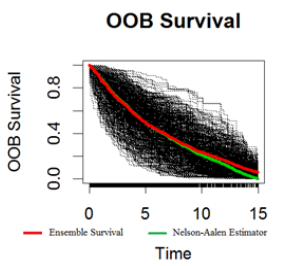
Table 30. Interactions Detection via VIMP ($\times 10^{-3}$) Analysis with Log-Rank Based RSF – Simulation Study

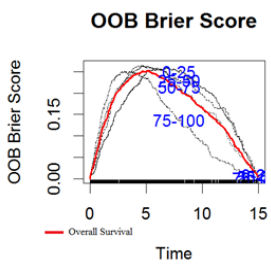
Interactions	Var	Var2	Paired	Additive	Diff
MAP:Sex	15.8	9.4	23.1	25.3	-2.1
MAP:Race	15.8	5.9	23.1	21.8	1.3
MAP:BMI	15.8	4.7	15.8	20.6	-4.8
MAP:Treatt	15.8	2.2	22.5	18.1	4.4
MAP:Age	15.8	1.2	17.8	17.0	0.8
Sex:Race	9.0	5.9	13.9	14.9	-1.0
Sex:BMI	9.0	4.7	12.1	13.7	-1.6
Sex:Treat	9.0	2.2	15.7	11.2	4.4
Sex:Age	9.0	1.2	9.8	10.2	-0.4
Race:BMI	5.3	4.7	8.8	10.0	-1.2
Race:Treat	5.3	2.2	11.8	7.5	4.3
Race:Age	5.3	1.2	6.0	6.4	-0.4
BMI:Treat	5.3	2.2	10.4	7.5	2.8
BMI:Age	5.3	1.2	4.1	6.5	-2.3
Age:Treat	2.6	1.2	9.4	3.8	5.6

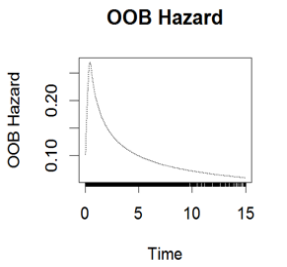
The cross validated log-rank based RSF model was then used for predictions of future events and assessment of prediction performance. The predicted survival probability, cumulative hazard and hazard function for a subset of 3 subjects are presented in Figure 47 and the overall OOB survival probability, OOB Brier scores and mortality are presented in Figure 48.











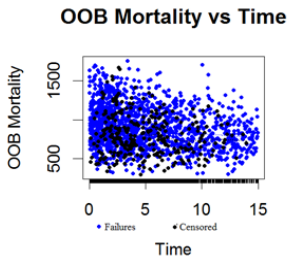


Figure 47. CV Survival, Cumulative Hazard and Hazard function for Log-Rank Based RSF (Subset of 3 Subjects) – Simulation Study

Figure 48. CV Survival, OOB Brier Scores and Mortality for Log-Rank Based RSF (All Subjects) – Simulation Study

The predicted mortality and predicted survival probability were obtained for the test set; the corresponding plots are presented in Figure 49 and Figure 50 respectively.

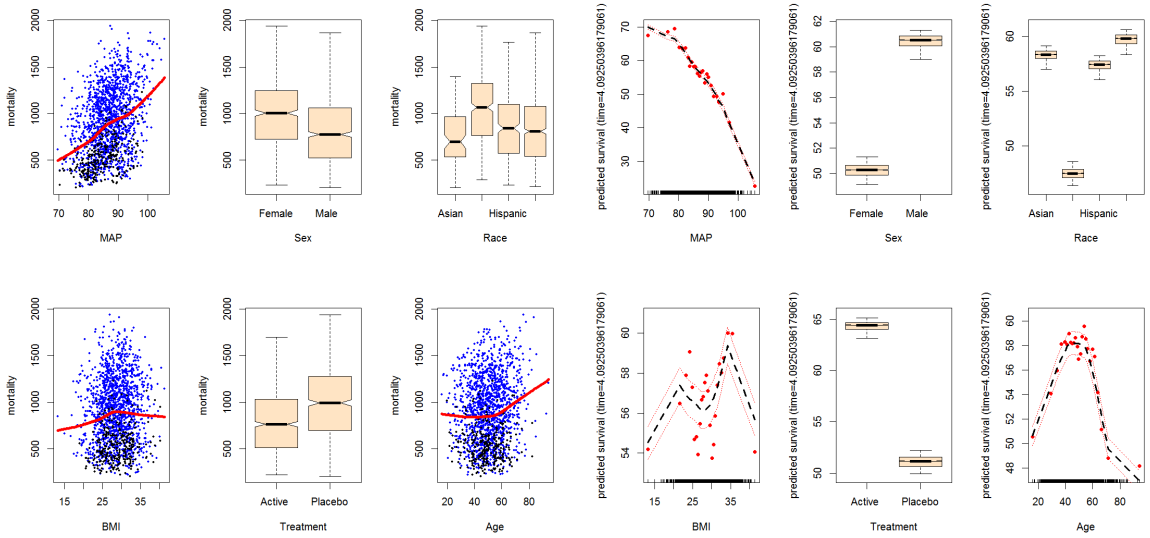


Figure 49. Predicted Mortality vs Each Factor from Log-Rank Based RSF – Simulation Study

Figure 50. Predicted Survival vs Each Factor from Log-Rank Based RSF – Simulation Study

4.1.2.3.3.2 Conditional Inference (CINF) Based RSF

Another nonparametric RSF model, conditional inference based random survival

forest (CINF-RSF), was also tuned with the training set via cross validation. A conditional inference based forest tree is presented in Figure 51; the survival plots are displayed in the terminal node.

The tree response from the cross validated conditional inference based RSF model was then used for prediction of future events or assessment of prediction performance based on the test set. Figure 52 presents the predicted survival probability for the test set based on the cross validated CIINF-RSF model, in which the Kaplan Meier curve in light green is the gold reference.

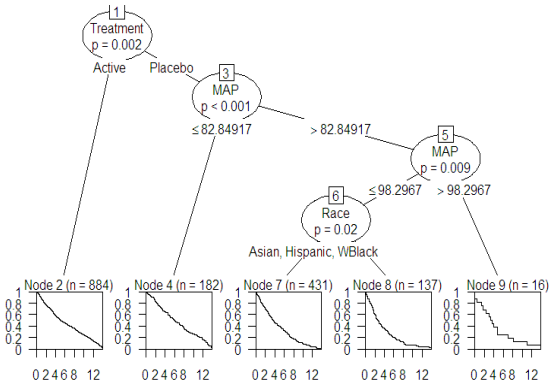


Figure 51. A Sample Forest Tree from Conditional Inference Based RSF – Simulation Study

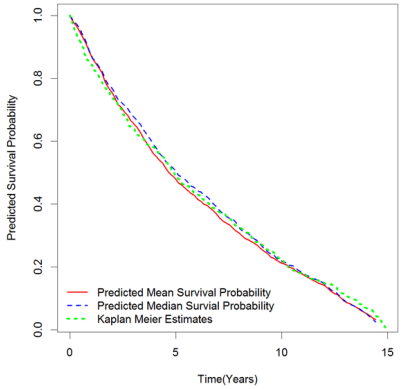


Figure 52. Prediced Survival from Conditional Inference Based RSF and Kaplan Meier Curve – Simulation Study

As discussed, both RSF models (LR-RSF and CINF-RSF models) managed to pick up the important prognostic factors, although slight difference was observed between the two models; next, the performance of the two models should be obtained in terms of prediction errors and time-dependent AUCs.

The prediction errors for log-rank based RSF (LR-RSF) and conditional inference based RSF (CINF-RSF) are summarized in Table 31; the corresponding 95 percentile credible intervals (PCI) were obtained with 200 bootstrap samples; unlike the other models, for which the 95% credible intervals were obtained from 1000 bootstrap samples, the two RSF models were built on bootstrap aggregation, it was very resource consuming to bootstrap 95% PCIs based on 1000 bootstrap sample (the process was attempted twice with 1000 bootstrap samples; each time, it ran out of memory after 96 hours of computer time); thus, 200 bootstrap samples were used instead to obtain the 95% PCIs, the

estimates were quite reasonable (see Table 31). For cross comparisons purpose, the prediction errors obtained from Cox PH linear model (as discussed in section 4.1.2.3.1) was used as reference. (The corresponding 95% PCIs for Cox PH linear model should have already been presented in Table 17, thus they are not again in Table 31.)

Table 31. Prediction Errors for Log-Rank Based RSF, Conditional Inference Based RSF and Conventional Cox Linear Model – Simulation Study Test Set

Yrs	LR-RSF (95% PCI)	CINF-RSF (95% PCI)	CoxLin
1	0.1109 (0.0909, 0.1325)	0.1123 (0.0929, 0.1337)	0.1118
2	0.1671 (0.1504, 0.1919)	0.1675 (0.1541, 0.1942)	0.1674
3	0.1986 (0.1838, 0.2176)	0.2023 (0.1912, 0.2187)	0.2021
4	0.2181 (0.2051, 0.2330)	0.2182 (0.2128, 0.2321)	0.2177
5	0.2425 (0.2290, 0.2549)	0.2330 (0.2298, 0.2495)	0.2324
6	0.2463 (0.2309, 0.2611)	0.2357 (0.2264, 0.2515)	0.2350
7	0.2490 (0.2309, 0.2654)	0.2339 (0.2242, 0.2557)	0.2330
8	0.2425 (0.2236, 0.2613)	0.2294 (0.2196, 0.2565)	0.2290
9	0.2302 (0.2081, 0.2527)	0.2219 (0.2068, 0.2507)	0.2214
10	0.2157 (0.1893, 0.2428)	0.2093 (0.1908, 0.2429)	0.2096
11	0.2021 (0.1753, 0.2303)	0.2000 (0.1778, 0.2335)	0.2006
12	0.1925 (0.1676, 0.2200)	0.1922 (0.1674, 0.2254)	0.1921
13	0.1853 (0.1524, 0.2134)	0.1879 (0.1590, 0.2254)	0.1877
14	0.1722 (0.1419, 0.2046)	0.1795 (0.1504, 0.2188)	0.1793

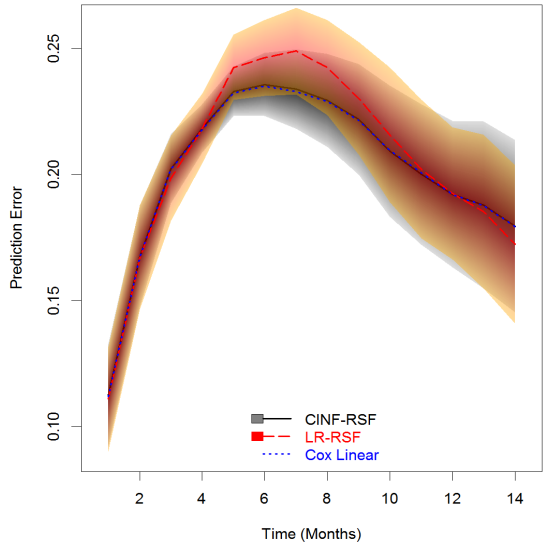


Figure 53. Prediction Errors for LR-RSF, CINF-RSF and Cox Linear Model – Simulation Study Test Set

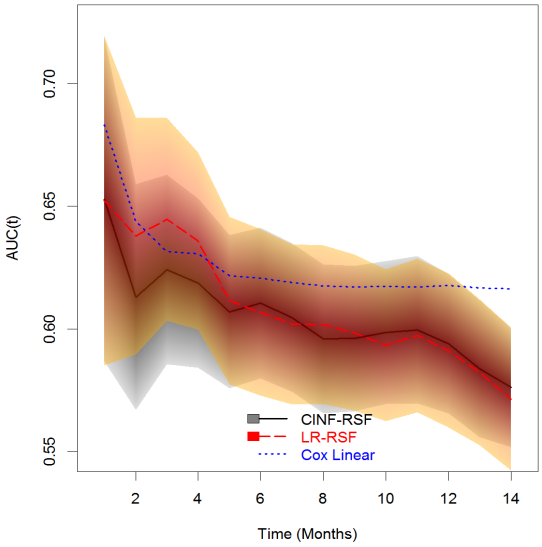


Figure 54. Prediction AUC for LR-RSF, CINF-RSF and Cox Regression – Simulation Study Test Set

The plots of the prediction errors for the two RSF models and the corresponding 95% PCIs are displayed in Figure 53. The black solid curve over the grey shaded area refers to the prediction errors and the corresponding 95% PCIs for CINF- RSF; the red dashed curve with the light pink shaded area is the prediction errors and the corresponding 95% PCIs for LR-RSF; the blue dotted curves is the reference prediction errors from the Cox PH linear model. In terms of prediction errors, the Cox PH linear model was slightly better than the two RSF models overall; the two RSF models were similar to each other within the first 5 years and the 95% PCIs were almost overlapping for this period. From 5 to 10 years, the log-rank based RSF model was slightly worse than the conditional inference based RSF model and at the tail (beyond year-10), the LR-RSF model was slightly better than the CINF-RSF model. Considering the nonparametric nature of the RSF models, the prediction errors were reasonably satisfactory.

Table 32. Prediction AUC for Log-Rank Based RSF, CINF Based RSF and Conventional Cox Model – Simulation Study Test Set

Yrs	LR-RSF (95% PCI)	CINF-RSF (95% PCI)	CoxLin
1	0.6523 (0.5838, 0.7181)	0.6529 (0.5887, 0.7211)	0.6832
2	0.6379 (0.5896, 0.6865)	0.6130 (0.5626, 0.6549)	0.6438
3	0.6447 (0.6028, 0.6857)	0.6242 (0.5845, 0.6618)	0.6316
4	0.6359 (0.6004, 0.6727)	0.6187 (0.5870, 0.6562)	0.6307
5	0.6115 (0.5803, 0.6485)	0.6070 (0.5763, 0.6388)	0.6218
6	0.6067 (0.5735, 0.6416)	0.6105 (0.5815, 0.6429)	0.6207
7	0.6019 (0.5704, 0.6358)	0.6046 (0.5764, 0.6370)	0.6190
8	0.6017 (0.5679, 0.6326)	0.5959 (0.5687, 0.6296)	0.6176
9	0.5984 (0.5676, 0.6314)	0.5961 (0.5695, 0.6289)	0.6171
10	0.5933 (0.5638, 0.6258)	0.5985 (0.5721, 0.6304)	0.6173
11	0.5973 (0.5677, 0.6306)	0.5996 (0.5719, 0.6320)	0.6172
12	0.5912 (0.5636, 0.6265)	0.5939 (0.5666, 0.6234)	0.6178
13	0.5824 (0.5541, 0.6137)	0.5837 (0.5580, 0.6140)	0.6167
14	0.5714 (0.5442, 0.6026)	0.5761 (0.5529, 0.6015)	0.6163

The time-dependent AUCs for the two RSF models as well as the 95% PCIs are presented in Table 32; the prediction AUCs from the Cox PH linear model are also presented as a reference (the 95% PCI for Cox PH linear model were presented in Table 17). The AUC curves of the three models are displayed in Figure 54. The black solid curve is the AUCs for conditional-inference based RSF model and the grey shaded area covers the 95% PCIs; the red dashed curve with the light orange shaded area is the AUCs and the 95% PCIs for LR-RSF model; the blue dotted curves is the reference AUCs from

the Cox PH linear model. Again, the Cox PH linear model was still the best of all 3 models in general and the two RSF models had the maximum difference within the first 5 years, after which the two RSF models had similar AUCs.

Comparing all three different models in terms of prediction performance and ease of use, the Cox PH linear model had the best prediction performance than the two RSF models. For ease of use, the two RSF models were efficient alternatives for assessing survival outcomes with reasonable performance; LR-RSF model is favorable if the covariates are not highly correlated, however the CINF-RSF model may be more reasonable for highly correlated data. Additionally, the two RSF models should both be able to deal with many more predictors than Cox PH linear model; considering the nonparametric nature of the two models, they were much easier to implement, since they had no model assumptions. However, the LR-RSF model could not accounted for multiple observations per subject caused by time-varying treatment effect, the multiple observations obtained from the same subject was considered as two independent subject, which was the only pitfalls for these LR-RSF model.

4.1.2.3.4 Penalized (Lasso, Ridge and Elastic-Net) Cox Regression Models

For penalized Cox regression models, three models were evaluated, lasso Cox regression, ridge Cox regression and elastic-net Cox regression. The R/glmnet package was implemented for assessment of the three penalized Cox regression models. Unfortunately, the penalized Cox regression models could not handle multiple observations per subject caused by time-varying treatment effect either, therefore the subject who switched from placebo to active treatment were considered as two independent subjects, one subject treated with placebo starting at day 0 and censored at the time of switching and another one treated with active treatment starting at the time of treatment switching, until the failures or censors. For this simulation study, this was reasonable since the treatment switching was independent of the failure event.

In terms of the formulation, the 3 models were very similar; the only difference was in the penalization terms; these approaches were originally developed to handle correlated high dimensional data, where the number of covariate were more than the total number of observations, i.e., $p \gg N$. Additionally, the penalized terms were introduced to regularize correlated factors. However, this simulation study only included a few

factors, thus correlated factors could be generated from polynomial transformations as well as interaction terms; the intention was to compare the performance of these penalized models to the typical Cox regression models, including generalized Cox linear model and the multivariate Cox regression models with nonlinear transformations.

Before implementing the models, all continuous variables were transformed to 5-degree polynomials terms, and the nominal variables were transformed into dummy binary variables and pairwise interactions were constructed between any two terms. A total of 345 polynomial and interaction covariate terms were constructed for the penalized Cox regression models; the 345 polynomial and interaction terms are presented in Appendix 1. As mentioned in section 3.3.2.1, lasso Cox regression could be cross validated by setting $\alpha = 1$ and ridge regression could be obtained by setting $\alpha = 0$, the partial log likelihood deviance of the lasso Cox regression and ridge Cox regression were displayed in Figure 55 and Figure 56, respectively.

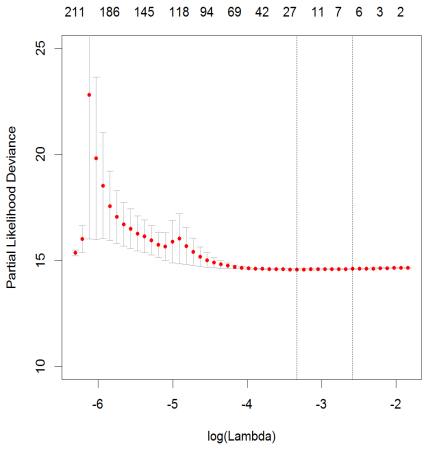


Figure 55. CV for Lasso Regression – Simulation Study

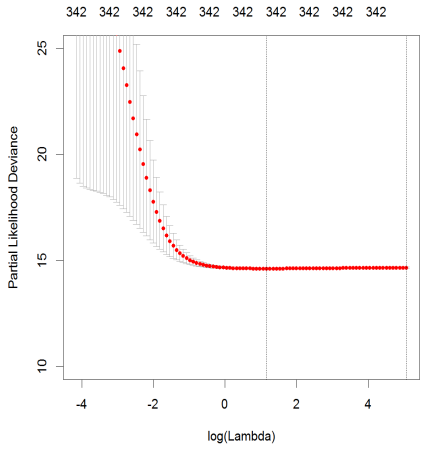


Figure 56. CV for Ridge Regression – Simulation Study

In the figures, the left vertical line colored in grey corresponds to the λ when the log partial likelihood deviance reaches the minimum and the right vertical line corresponds to the λ for the regularized model with deviance within 1 standard deviation of the minimum. The numbers above the figure indicates the number of terms left in the model. For lasso Cox regression, the model reached the minimum log partial likelihood deviance at λ of 0.0357, where the model kept 23 terms in the model; for ridge regression, the model reached the minimum deviance at λ of 3.1815, where the model included 342 terms in the model. For the selected "best" models, the corresponding regression

coefficients for lasso and ridge Cox regression are presented in Appendix 2 and Appendix 3, respectively. However, it was noted that the results from the lasso and ridge regression were not very stable; for a different seed, the results could be slightly different; such unreliability would potentially limit the generalization of the approaches, and it would be impossible to make accurate predictions based on the selected "best" models.

To improve the robustness, a different cross validation process was developed to achieve more stable results which did not change over different processes or different seeds. The idea was to use Brier score as the selection rule for cross validate the lasso and ridge Cox models, where the empirical time-dependent Brier scores, $BS(t)$, were calculated using the formula from **Error! Reference source not found.**. The goal was to select the penalization term, λ and the corresponding penalized models which could achieve a minimum Brier score.

The lasso Cox model achieved the minimum Brier scores with λ = 0.2855. The CV errors (Brier scores) are displayed in Figure 57. With this process, different seeds were attempted for the CV; results were quite stable, the selected λ and the lasso Cox model corresponding to the λ from the CV did not change with the different processes or seeds.

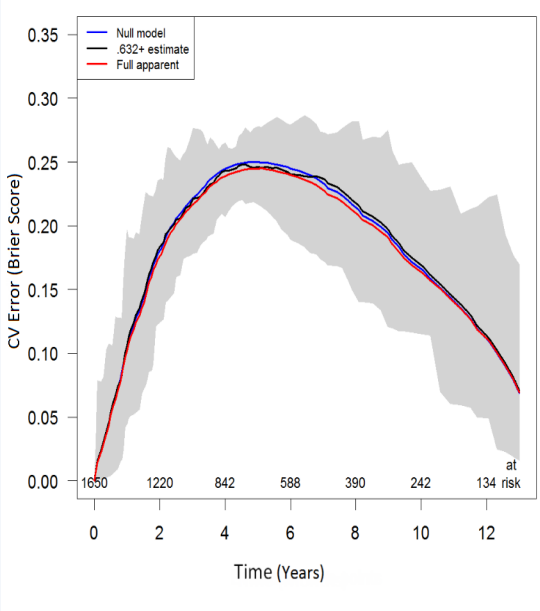


Figure 57. CV Error for Lasso Cox Regression with λ = 0.2855 – Simulation Study

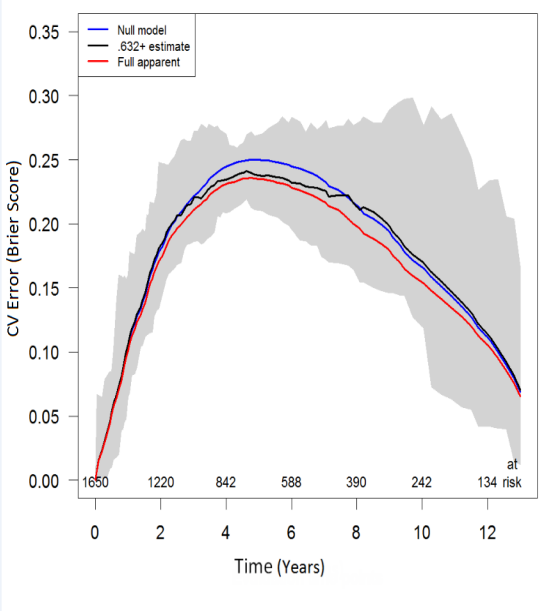


Figure 58. CV Error for Ridge Cox Regression with λ = 1.8153 – Simulation Study

Similarly the best ridge regression model was selected using cross validation via Brier's score. The ridge regression reached its best performance with $\lambda = 1.8153$; the cross validation Brier's scores are displayed in Figure 58. In the two figures (Figure 57 and Figure 58), the blue solid line is cross validation error of the null model in which the survival probability was estimated with the Cox model without any covariates over the entire training set and the red solid line is the CV error of the full model with penalization parameters obtained from the entire training set; the black solid line is the CV error of the selected "best" model, in which the survival probability was estimated from lasso (or ridge) Cox regression over the 10^{th} (left-out) CV samples and adjusted with .632 rule as suggested by Efron et al. (1997)[151] based on the best model with the penalization parameters obtained from 9 CV samples; the gray shaded area is covered by the resampling data.

With $\lambda = 0.2855$, a total of 10 covariate terms were kept by the lasso Cox model and the corresponding coefficients for the "best" lasso Cox regression model are presented below. The regression coefficients were kept 5 decimal places, in order to retain the interaction term of $\text{Age}^5\text{: Race} = \text{White}$, which is presented as $\text{Age}^5\text{: \{White\}}$.

$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}$, where

$$X\hat{\beta} = 0.00327\ \text{MAP}^4 + 0.00650\ \text{MAP}^5 + 0.00298\ \text{Age}^4\text{: \{Placebo\}} - 0.00018\ \text{Age}^5\text{: \{Male\}} - 0.00004\ \text{Age}^5\text{: \{White\}} + 0.00667\ \text{MAP}^5\text{: \{Placebo\}} + 0.00178\ \text{MAP}^5\text{: \{Male\}} - 0.00378\ \text{BMI}^4\text{: \{Male\}} - 0.00104\ \text{BMI}^5\text{: \{Placebo\}} - 0.00094\ \text{BMI}^5\text{: \{Hispanic\}}$$

Similarly, with λ=1.8153, the ridge Cox regression achieved the optimum CV Brier Scores with a total of 342 covariate terms; the coefficients for the 342 covariate terms are not presented in this paper, since they were only slightly different from the ones cross validated via the minimum partial log likelihood deviance (see Appendix 3 for the coefficients for ridge Cox model from CV with partial log likelihood deviance).

For elastic-net Cox regression, both α and λ should be needed. With any given α, the λ could be searched automatically; however, the search of α was quite computation intensive, theoretically the alpha could be set as an entry of the sequence from 0.1 to 0.9 with a step of 0.1, then use CV to search for the λ corresponding to the best performance in terms of partial log likelihood deviance, then the number of covariate terms were

obtained accordingly. The goal of this step was to achieve the minimum number of terms. Such approach was named as exhaustive searching (ES). Figure 59 presents the exhaustive search paths. The best partial log likelihood deviance was not monotone with α; as can be seen from the figure, the elastic-net Cox regression model with alpha=0.68, achieved the minimum deviance with only 12 covariate terms based on the selected sequence. After the "best" α (0.68) was identified, the penalization parameter, λ, was searched again for the minimum partial log likelihood deviance, λ =0.0695 should achieve the minimum deviance; the search paths for λ with α set to 0.68 are displayed in Figure 60.

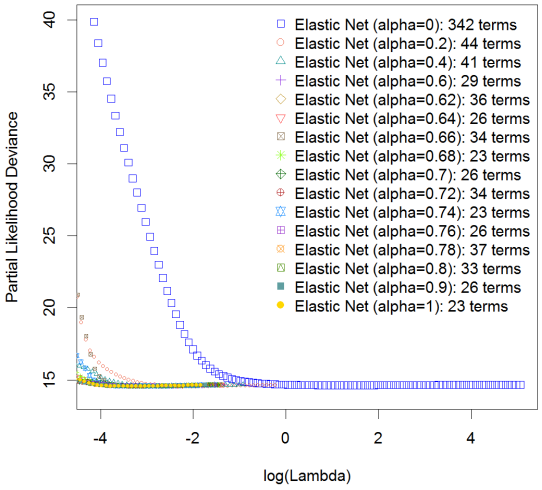


Figure 59. CV for Selection of α for Elastic-Net Cox Regression – Simulation Study

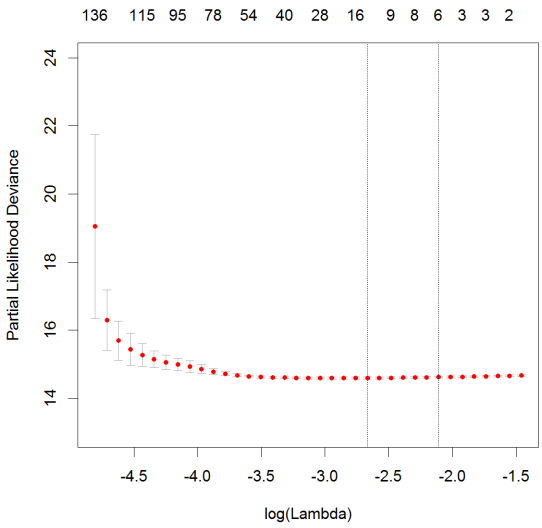


Figure 60. CV for Elastic-Net Cox Regression with α=0.68 – Simulation Study

The elastic-net Cox regression model with α=0.68 and λ = 0.0695 as obtained from exhaustive search algorithm retained 12 covariate terms; the corresponding model is formulated as

$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}$, where

$$X\hat{\beta} = 0.0914\ \text{MAP} + 0.0011\ \text{Age:MAP}^2 + 0.0064\ \text{MAP}^3 + 0.0014\ \text{MAP}^5 + 0.0002\ \text{BMI}^2 + 0.0155\ \text{Age:MAP} - 0.3092\ \{\text{Active}\} - 0.0956\ \{\text{Male}\} + 0.0028\ \text{Age}^3\text{:MAP} - 0.0242\ \text{BMI}^2\text{:}\{\text{Male}\} - 0.0083\ \text{Age}^2\text{:BMI}^2 - 0.0477\ \{\text{Male}\}\text{:}\{\text{White}\}$$

And, the coefficients estimated from the cross validated elastic-net Cox model are presented in Table 33.

Table 33. Coefficients of the Selected Elastic-Net Cox Regression Model with Penalization Parameters Obtained from Exhaustive Search (α =0.68 and λ =0.0695) – Simulation Study

Terms	Coefs	Terms	Coefs
MAP	0.0914	Sex=Male	-0.0956
Age:MAP^2	0.0011	Age:MAP:BMI:Sex=Male	0.0107
MAP^3	0.0064	Age^3:MAP:BMI:Race=Hispanic	0.0028
MAP^5	0.0014	BMI^2:Sex=Male	-0.0242
Age:MAP:BMI^2	0.0155	Age^2:BMI^2:Race=Asian	-0.0083
Treatment=Active Treatment	-0.3092	Sex=Male:Race=White	-0.0477

The partial log likelihood deviance of the three models, ridge, lasso and elastic-net Cox regression are compared in Figure 61, in which, the penalization parameters, λ's for the lasso and ridge Cox models, were obtained from the CV using Brier score as the selection rule.

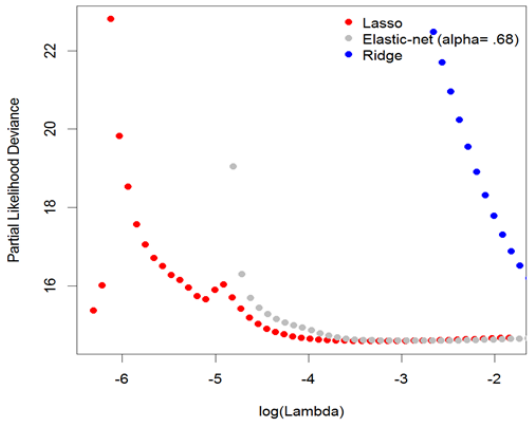


Figure 61. CV Performance of Ridge, Lasso and Elastic-Net Cox models – Simulation Study

However for elastic-net Cox regression, the penalization parameters (α and λ) obtained from the exhaustive search were still extremely unstable, different seeds might yield different penalization parameters and the training process was very time consuming; more importantly, there was no guarantee of global minimum of partial log likelihood deviance or cross validation errors; further, the search of alpha was very much depending on the searching sequence; i.e., the step of 0.1 for the selected sequence from 0.1 to 0.9 could be too big to find a minimum. In fact, any pre-specified step would not be a good idea, since the partial log likelihood deviance was not monotonic over the range of penalization parameters and the minimum of the deviance did not correspond to the minimum of CV errors.

With the above considerations, an interval search algorithm (IS) was developed, which was originally proposed for support vector machine learning. The idea was to globally search both penalization parameters simultaneously through a Gaussian model of the error surface in the parameter space and sampling systematically towards the global minimum of the Brier scores. The algorithm still did not guarantee to find the global minimum of the partial log likelihood deviance, but it could achieve the minimum cross validation errors (Brier scores). Additionally, the search path was very efficient; both penalization terms (α and λ) were searched simultaneously. Furthermore, the penalization parameters obtained from the interval search were quite reliable; different seeds were attempted again for the searching process, the same penalization parameters were achieved and the corresponding model also achieved the same Brier scores and partial log likelihood deviance.

The interval search paths for penalization parameters are displayed in Figure 62; the partial log likelihood deviances are presented on the top border of the figure; the penalization parameters are presented on the x- and y-axis and the number of covariate terms is labelled next to each point. The coordinates to the intersection of the red solid lines reflect the penalization terms when the elastic-net Cox regression model achieved the optimum CV performance, where α =0.2321 and λ = 0.2234.

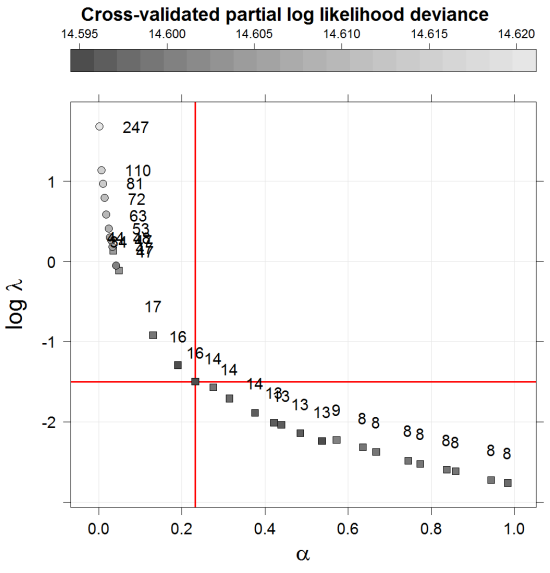


Figure 62. Interval Search Paths for Elastic Net Cox Regression – Simulation Study

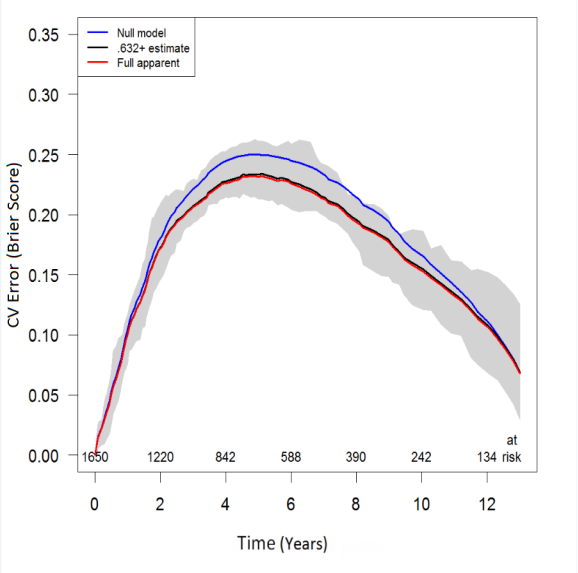


Figure 63. CV Brier Score for Elastic Net Cox Regression with α =0.2321 and λ = 0.2234 – Simulation Study

The CV errors of the corresponding elastic-net Cox model are displayed in Figure 63. It can be seen from the figure that the elastic-net Cox model had the almost the same CV errors (black solid curve) as the full apparent model (red solid curve), which suggested that the elastic-net Cox regression had achieved almost the same amount of information as the full apparent model with respect to the survival outcome.

With penalization terms α =0.2321 and λ = 0.2234, the model retained a total of 16 covariate terms; the coefficients of the 16 covariate terms for the best elastic-net Cox regression model are summarized in Table 34; the corresponding elastic-net Cox regression model was formulated with the coefficients from the table.

Table 34. Coefficients of the Best Elastic-Net Cox Model with Penalization Parameters (α =0.2321 and λ = 0.2234) from Interval Search – Simulation Study

Terms	Coeffs
MAP	0.0482
MAP^3	0.0078
MAP^5	0.0008
Age:MAP:BMI^2	0.0039
Treatment=Placebo	0.2457
Sex=Male	-0.0794
Age^2:Treatment=Placebo	0.0089
MAP:Treatment=Placebo	0.0629
MAP^3:Treatment=Placebo	0.0063
MAP^5:Sex=Male	0.0001
Age:MAP:BMI:Treatment=Placebo	0.0112
Age^3:MAP:BMI:Race=Hispanic	0.0018
BMI^2:Sex=Male	-0.0200
Age^2:BMI^2:Race=Asian	-0.0046
Age:MAP:BMI^2:Treatment=Placebo	0.0165
Sex=Male:Race=White	-0.0387

$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \text{where}$$
$$X\hat{\beta} = 0.0482\ \text{MAP} + 0.0078\ \text{MAP}^3 + 8 \times 10^{-4}\ \text{MAP}^5 + 0.0039\ \text{Age:MAP:BMI}^2 + 0.2457\ \{\text{Placebo}\} - 0.0794\ \{\text{Male}\} + 0.0089\ \text{Age}^2\text{:}\{\text{Placebo}\} + 0.0629\ \text{MAP:}\{\text{Placebo}\} + 0.0063\ \text{MAP}^3\text{:}\{\text{Placebo}\} + 10^{-04}\ \text{MAP}^5\text{:}\{\text{Male}\} + 0.0112\ \text{Age:MAP:BMI:}\{\text{Placebo}\} + 0.0018\ \text{Age}^3\text{:MAP:BMI:}\{\text{Hispanic}\} 0.0200\ \text{BMI}^2\text{:}\{\text{Male}\} - 0.0046\ \text{Age}^2\text{:BMI}^2\text{:}\{\text{Asian}\} + 0.0165\ \text{Age:MAP:BMI}^2\text{:Placebo} + -0.0387\ \{\text{Male}\}\text{:}\{\text{White"}\}$$

Table 35 compares the CV Brier scores across lasso, ridge Cox models, elastic-net Cox model via interval search and the elastic-net Cox model via exhaustive search; the plots of CV errors against time are also displayed in Figure 64. The elastic-net Cox model via exhaustive search and interval search were almost overlaid on top of each other; the two models had almost the same prediction errors (only slight differences were observed). Please note that for cross validation errors (Brier scores), the 95% PCIs could not be obtained from 10-fold CV (to obtain the 95% PCIs, a bootstrap process with at

least 100-fold should be needed). The two elastic-net Cox models with different searching algorithms had achieved almost the same CV errors; the elastic-net Cox model via exhaustive search was not as reliable due to the unstable penalization parameters obtained from the CV searching processes; since the different penalization terms could have resulted in selection of different covariates; as a results, it was difficult to generalize the application for broader use. While the selected elastic-net Cox regression model obtained from the interval search algorithm should have overcome these disadvantages, thus it was further assessed for the prediction performance.

Table 35. CV Errors (Brier Score) for Elastic Net, Lasso and Ridge Cox Regression – Simulation Study Training Set

Yrs	Elastic-Net (IS)	Elastic-Net (ES)	Lasso	Ridge
1	0.1013	0.1009	0.1063	0.1068
2	0.1750	0.1742	0.1851	0.1867
3	0.2085	0.2075	0.2212	0.2198
4	0.2280	0.2273	0.2430	0.2318
5	0.2337	0.2329	0.2460	0.2339
6	0.2277	0.2269	0.2403	0.2281
7	0.2146	0.2139	0.2362	0.2212
8	0.1958	0.1954	0.2179	0.2078
9	0.1794	0.1792	0.1975	0.2004
10	0.1543	0.1542	0.1683	0.1704
11	0.1330	0.1331	0.1424	0.1439
12	0.1069	0.1069	0.1120	0.1128
13	0.0686	0.0687	0.0706	0.0705

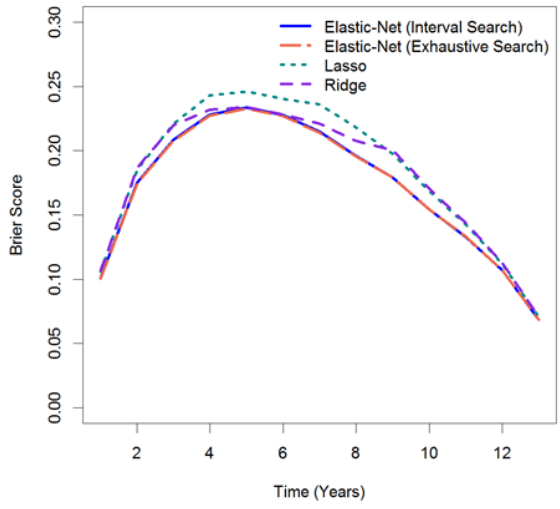


Figure 64. CV Errors (Brier Score) for Elastic Net, Lasso and Ridge Cox Regression (Training Set) – Simulation Study

For performance comparisons, the prediction performance measurements of the three selected penalized Cox models (lasso, ridge and elastic-net Cox regression models via interval search) were assessed based on the test set. The prediction errors (Brier scores) and time-dependent AUCs of the selected penalized Cox regression models are summarized in Table 36; the corresponding 95% PCIs for prediction errors and time-dependent AUCs obtained from 1000 bootstrap samples are also presented.

Table 36. Prediction Errors and Time Dependent AUCs for Lasso, Ridge and Elastic-Net Cox Models – Simulation Study Test Set

	Years	Lasso (95% PCI)	Ridge (95% PCI)	Elastic-Net IS (95% PCI)
Prediction Errors	1	0.113 (0.093, 0.137)	0.115 (0.093, 0.141)	0.115 (0.096, 0.138)
	2	0.175 (0.157, 0.193)	0.175 (0.155, 0.196)	0.173 (0.157, 0.190)
	3	0.213 (0.200, 0.226)	0.213 (0.196, 0.229)	0.209 (0.197, 0.222)
	4	0.230 (0.222, 0.237)	0.227 (0.215, 0.238)	0.225 (0.218, 0.233)
	5	0.243 (0.238, 0.247)	0.241 (0.233, 0.251)	0.238 (0.232, 0.243)
	6	0.244 (0.236, 0.251)	0.242 (0.237, 0.251)	0.239 (0.230, 0.24)
	7	0.242 (0.230, 0.252)	0.239 (0.243, 0.251)	0.237 (0.224, 0.249)
	8	0.234 (0.219, 0.248)	0.233 (0.216, 0.248)	0.233 (0.216, 0.247)
	9	0.227 (0.206, 0.246)	0.227 (0.205, 0.246)	0.227 (0.206, 0.247)
	10	0.213 (0.188, 0.236)	0.218 (0.192, 0.242)	0.214 (0.189, 0.240)
	11	0.206 (0.176, 0.236)	0.209 (0.180, 0.238)	0.208 (0.178, 0.239)
	12	0.199 (0.171, 0.236)	0.203 (0.175, 0.237)	0.201 (0.172, 0.239)
	13	0.193 (0.159, 0.231)	0.196 (0.165, 0.236)	0.196 (0.163, 0.235)
	14	0.180 (0.144, 0.221)	0.182 (0.150, 0.226)	0.184 (0.149, 0.229)
Time Dependent AUCs	1	0.644 (0.584, 0.715)	0.666 (0.604, 0.728)	0.690 (0.631, 0.750)
	2	0.599 (0.553, 0.649)	0.634 (0.583, 0.681)	0.649 (0.603, 0.697)
	3	0.590 (0.548, 0.632)	0.627 (0.584, 0.670)	0.636 (0.591, 0.677)
	4	0.590 (0.550, 0.628)	0.625 (0.584, 0.662)	0.634 (0.593, 0.672)
	5	0.586 (0.549, 0.623)	0.616 (0.578, 0.652)	0.626 (0.590, 0.662)
	6	0.582 (0.549, 0.620)	0.614 (0.579, 0.649)	0.625 (0.591, 0.661)
	7	0.579 (0.545, 0.616)	0.614 (0.579, 0.647)	0.625 (0.591, 0.657)
	8	0.576 (0.542, 0.611)	0.611 (0.576, 0.643)	0.621 (0.586, 0.651)
	9	0.575 (0.542, 0.609)	0.609 (0.575, 0.640)	0.619 (0.585, 0.650)
	10	0.574 (0.541, 0.607)	0.606 (0.573, 0.636)	0.618 (0.584, 0.648)
	11	0.574 (0.542, 0.608)	0.607 (0.574, 0.636)	0.612 (0.584, 0.647)
	12	0.574 (0.542, 0.607)	0.606 (0.573, 0.636)	0.618 (0.585, 0.647)
	13	0.573 (0.542, 0.607)	0.606 (0.573, 0.635)	0.617 (0.584, 0.647)
	14	0.573 (0.542, 0.606)	0.605 (0.573, 0.635)	0.617 (0.584, 0.647)

The graphic display of the prediction errors, the time-dependent AUCs and the corresponding 95% PCIs are presented in Figure 65. The solid black line and the grey shaded area indicate the prediction errors (or AUCs) and the corresponding the 95% PCIs for the lasso Cox model; the dotted blue line and the light blue shaded area are the

prediction errors and the corresponding 95% PCIs for ridge Cox regression model; the prediction errors for elastic-net was the orange line covered by the orange shaded area for the corresponding 95% PCIs. In terms of the prediction errors, the three models were very similar to each other. As can be seen from the figure, in terms of the time-dependent AUCs, the elastic-net had the best performance and the lasso Cox model had the worst performance.

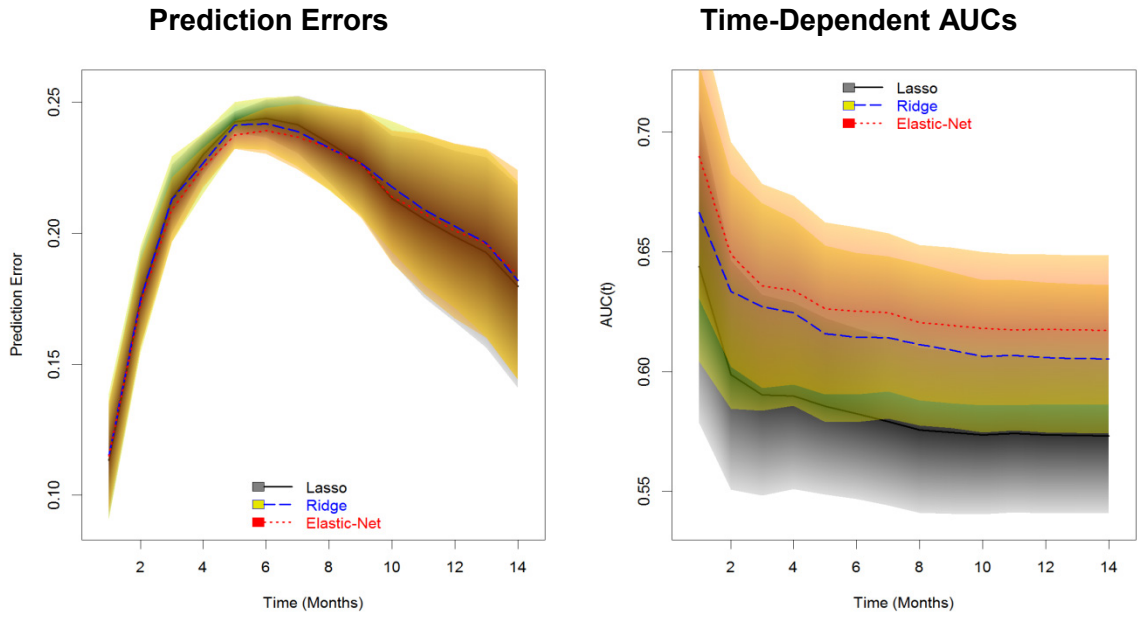


Figure 65. Prediction Errors and Time-Dependent AUCs for Lasso, Ridge and Elastic-Net Cox Models – Simulation Study Test Set

Additionally, comparing with the typical Cox regression models, the regression coefficients from the penalized Cox regression models including ridge, lasso or elastic-net Cox regression were all biased, because of the regularization from the penalization parameter(s). Unbiased estimates of the regression coefficients were obtained by fitting typical Cox regression models with the exact same covariate terms as obtained from the above penalized Cox model. The unbiased estimates of the coefficients for lasso and elastic-net Cox models are reported in Table 37 and Table 38. For ridge Cox regression, the model was non-estimable due to too many covariate terms (342). On the other hand, the ridge Cox regression model was able to achieve good performance only by keeping the most (if not all) of the covariate terms; therefore even if there were enough data, it was still not possible to use ridge Cox model to perform prognostic factor selection.

Table 37. Unbiased Coefficients for Lasso Cox Model – Simulation Study

	Coef	HR	SE (Coef)	z	P-value
MAP^4	0.0104	1.0105	0.0039	2.64	0.0082
MAP^5	0.0051	1.0051	0.0028	1.85	0.0648
Age^4:Treatment=Placebo	0.0113	1.0113	0.0032	3.47	0.0005
Age^5:Sex=Male	-0.0019	0.9981	0.0010	-1.85	0.0648
Age^5:Race=White	-0.0017	0.9983	0.0011	-1.49	0.1350
MAP^5:Treatment=Placebo	0.0086	1.0087	0.0031	2.79	0.0053
MAP^5:Sex=Male	0.0040	1.0040	0.0032	1.27	0.2035
BMI^4:Sex=Male	-0.0117	0.9883	0.0039	-3.01	0.0026
BMI^5:Treatment=Placebo	-0.0026	0.9974	0.0016	-1.66	0.0975
BMI^5:Race=Hispanic	-0.0033	0.9967	0.0014	-2.27	0.0232

Table 38. Unbiased Coefficients for Elastic-Net Cox Model (IS) – Simulation Study

	Coef	HR	SE (Coef)	z	P-value
MAP	0.2211	1.2474	0.0813	2.72	0.0066
MAP^3	-0.1139	0.8923	0.0490	-2.32	0.0201
MAP^5	0.0184	1.0186	0.0075	2.46	0.0141
Age:MAP:BMI^2	0.0187	1.0188	0.0232	0.81	0.4205
{Placebo}	0.4256	1.5306	0.0617	6.90	0.0000
Sex=Male	-0.1499	0.8608	0.0720	-2.08	0.0374
Age^2:{Placebo}	0.0290	1.0294	0.0255	1.14	0.2553
MAP:{Placebo}	0.0656	1.0678	0.0978	0.67	0.5023
MAP^3:{Placebo}	0.0437	1.0447	0.0320	1.37	0.1712
MAP^5:{Male}	0.0021	1.0021	0.0038	0.55	0.5829
Age:MAP:BMI:{Placebo}	0.0801	1.0834	0.0424	1.89	0.0585
Age^3:MAP:BMI:{Hispanic}	0.0418	1.0427	0.0209	2.00	0.0460
BMI^2:{Male}	-0.0467	1.2474	0.0252	-1.85	0.0641
Age^2:BMI^2:{Asian}	-0.0756	0.8923	0.0646	-1.17	0.2420
Age:MAP:BMI^2:{Placebo}	0.0605	1.0186	0.0323	1.87	0.0610
Sex=Male:{White}	-0.1496	1.0188	0.0705	-2.12	0.0339

HR = exp(Coef);

4.1.2.3.5 Principal Component Cox Regression (PCR)

Principal component Cox regression (PCR) model was implemented and cross validated; the multiple observations due to subject switching treatment were adjusted with AG counting process extension. Note that the initial tentative fit of principal component regression was already discussed in section 4.1.2.2.3; the AIC performance of the PCR model was not very good. Therefore, in this section, the analysis results were only be briefly discussed and summarized.

It can be seen from Figure 4 that the first 6 components contributed to the 99% of the total variance and the Figure 5 provided additional evidence that component 7 did not improve the performance of the PCR model. Therefore, the initial PCR model only included the first 6 components.

The 6-component PCR model was further selected via a backward step-down procedure based on AIC criterion; again the AIC was preset to ≥1e-10 to ensure all components to be deleted from the model. Figure 66 presents the deleted components vs. the AIC and dfs remained for the PCR model. The AIC reaches the minimum after component 4 was deleted, therefore the final model should include the rest of the 5 components.

Deleted Components	df remained	AIC
None	6	17474.6
-Comp.4	5	17473.7
-Comp.6	4	17474.0
-Comp.3	3	17477.0
-Comp.5	2	17517.5
-Comp.1	1	17561.2
-Comp.2	0	17577.1

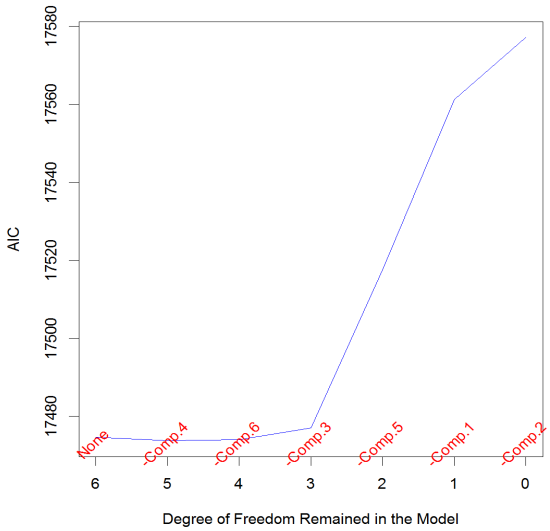


Figure 66. AIC vs. df. of the Remaining PCR Models after Each Component Deletion – Simulation Study

The coefficients of the 5-component PCR model are presented in Table 39; all of the components were significant, with p-value < 0.05.

Table 39. Coefficients of the 5-Component PCR Model – Simulation Study

	Coef	HR	SE (coef)	z	Pr(>|z|)
Comp.1	0.1412	1.1517	0.0198	7.12	0.0000
Comp.2	0.2072	1.2302	0.0271	7.65	0.0000
Comp.3	0.0585	1.0602	0.0268	2.18	0.0291
Comp.5	0.1918	1.2114	0.0289	6.63	0.0000
Comp.6	0.0399	1.0407	0.0267	1.49	0.1360

The model selected from the above step were evaluated via 10-fold leave-one-out cross validation, the performance statistics are presented in Table 40. There was no

concern of overfitting or underfitting.

Table 40. Cross Validation Performance for Principal Component Cox Regression – Simulation Study

	index.orig	training	test	optimism	index.corrected
Dxy	-0.1855	-0.1861	-0.1810	-0.0050	-0.1804
R2	0.0824	0.0828	0.0804	0.0024	0.0800
Slope	1.0000	1.0000	0.9830	0.0170	0.9830
D	0.0080	0.0082	0.0114	-0.0032	0.0112
U	-0.0001	-0.0001	0.0008	-0.0009	0.0008
Q	0.0081	0.0083	0.0106	-0.0023	0.0104
g	0.4054	0.4063	0.3978	0.0086	0.3968

Prediction performance of the selected PCR model was also assessed with the test set, including prediction errors and time-dependent AUCs. Table 41 presents the estimated prediction errors and time-dependent AUCs for the selected PCR model based on the test set. The corresponding plots are presented in Figure 67.

The coefficients of the components from the PCR model had to be converted back to the original factor for interpretations; for this reason, the loading matrix of the selected PCR model would be helpful, which should indicate the contribution of each factor to the total variance of the model; however it was still quite difficult to intuitively interpret the results since the survival outcomes were indirectly linked to the original factor via the latent components.

Table 41. Prediction Errors and Time-Dependent AUCs for PCR – Simulation Study Test Set

Yrs	Pred Errors (95% PCI)	AUCs (95% PCI)
1	0.1132 (0.0955, 0.1321)	0.6621 (0.5985, 0.7198)
2	0.1696 (0.1518, 0.1866)	0.6329 (0.5840, 0.6803)
3	0.2064 (0.1924, 0.2190)	0.6159 (0.5715, 0.6567)
4	0.2186 (0.2083, 0.2292)	0.6199 (0.5808, 0.6569)
5	0.2332 (0.2227, 0.2441)	0.6136 (0.5781, 0.6460)
6	0.2354 (0.2231, 0.2481)	0.6109 (0.5780, 0.6422)
7	0.2327 (0.2182, 0.2478)	0.6119 (0.5797, 0.6411)
8	0.2278 (0.2107, 0.2463)	0.6104 (0.5801, 0.6394)
9	0.2218 (0.2015, 0.2448)	0.6089 (0.5783, 0.6367)
10	0.2098 (0.1852, 0.2368)	0.6086 (0.5787, 0.6363)
11	0.2009 (0.1731, 0.2294)	0.6086 (0.5789, 0.6358)
12	0.1912 (0.1622, 0.2253)	0.6093 (0.5792, 0.6365)
13	0.1835 (0.1509, 0.2194)	0.6087 (0.5792, 0.6356)
14	0.1739 (0.1392, 0.2130)	0.6084 (0.5789, 0.6351)

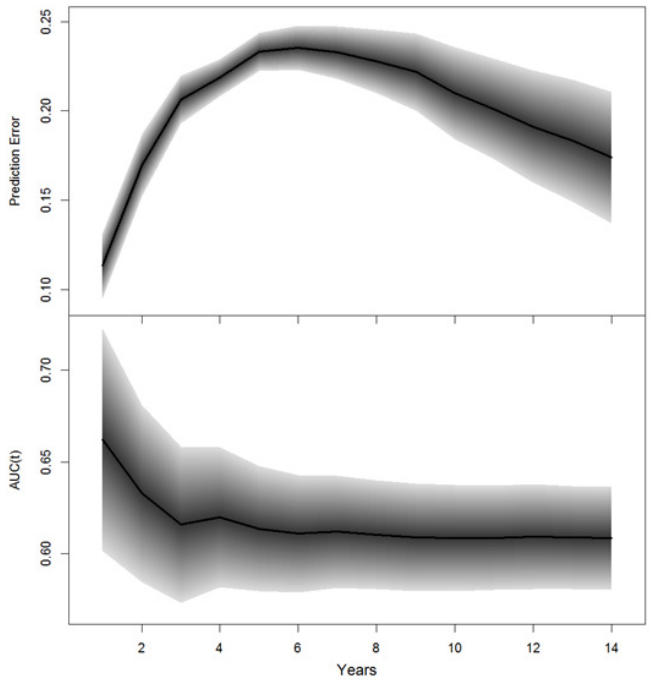


Figure 67. Prediction Errors and Time-Dependent AUCs for PCR – Simulation Study Test Set

4.1.2.3.6 Partial Least Squares Cox Regression

Partial least squares regression models have been reported to fit continuous and categorical outcomes with success; but very little has been published on survival outcomes. The same approach was developed on top of Cox PH model for evaluation of survival outcomes; several features and analyses tools were developed to assess predictions and prediction performance. The partial least squares Cox (PLS-Cox) model shared some resemblance with PCR model; latent components were initially derived to achieve the maximum correlation with respect to the survival outcome; the PLS components would be used as covariates for Cox regression model.

The PLS components were constructed with all factors in their original scale, including all correlated factors; once constructed, the components should be orthogonal (uncorrelated) to each other but pointing in the direction of the log hazard. Therefore, multicollinearity should be of no concerns. In addition, as Ron Wehrens (2011)[152] pointed out that polynomial transformations of the covariates generally should not improve the model performance, thus in this simulation study, all factors in their original

scale were attempted for the PLS Cox model. Multiple observations with time-varying treatment effect were adjusted using AG extension.

The PLS Cox model including all factors in their original scale and all pair-wise interactions, was cross validated with the training set. A total of 52 covariate terms were initially included in the model (see Appendix 4 for complete list of terms). The cross validation performance, model AIC corresponding to the number of components retained in the model is presented in Figure 68. The AIC curve suggested that the PLS Cox model reached the best performance with the first 8 components.

Comps	AIC	Comps	AIC
0	17573.9	8	17412.5
1	17462.4	9	17414.3
2	17460.3	10	17415.8
3	17450.6	11	17416.6
4	17452.1	12	17418.6
5	17435.6	13	17420.6
6	17436.8	14	17417.2
7	17424.7	15	17418.1

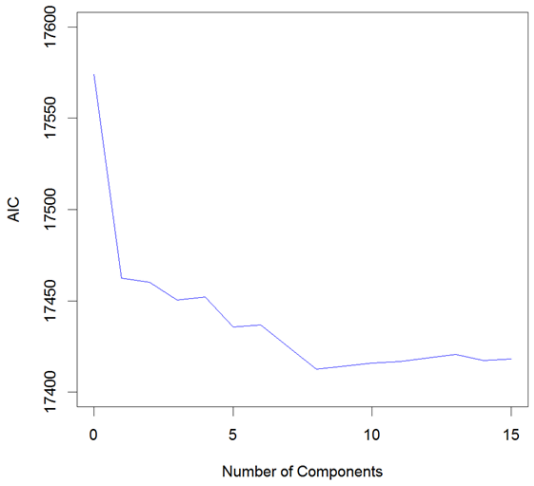


Figure 68. Model Performance of PLS Cox Linear Model – Simulation Study

Next, the 8-component PLS model were evaluated over the training set; the regression coefficients of the 8 components were estimated from a typical Cox regression model, which are presented in Table 42. For this model, the order of the components should be very important, the latter components were derived based on the results of earlier components; i.e., components 2 were not significant (p-value =0.62), but it had to be kept in the final model, since it was needed for deriving component 3 and above. Therefore, if a particular component was selected by the cross validation process, all components prior to that should also be included in the final model no matter whether they were significant or not.

Additionally, the original factors were linked to the log hazard through the latent components, therefore the coefficients for the PLS components had to be converted back to the original factors for interpretation (see Appendix 5 for the coefficients of the

original covariate terms as converted from the PLS components).

Table 42. Regression Coefficients from PLS Cox Regression Model with Linear Terms (1st Degree Polynomials) – Simulation Study

Comp	Coef	HR	SE (Coef)	z	p
#1	-0.1243	0.8831	0.0110	-11.33	<.0001
#2	-0.0256	0.9747	0.0119	-2.15	0.0315
#3	-0.0563	0.9452	0.0138	-4.08	<.0001
#4	0.0105	1.0105	0.0154	0.68	0.4944
#5	0.0839	1.0875	0.0194	4.33	<.0001
#6	-0.0169	0.9833	0.0159	-1.06	0.2882
#7	0.0655	1.0677	0.0170	3.85	0.0001
#8	-0.0978	0.9068	0.0257	-3.80	0.0001

Then the 8-component Cox PH model can be validated using 10-fold cross validation; results are presented in Table 43. There was no indication of underfitting or overfitting; however, these statistics were obtained from the typical Cox regression model with the 8 PLS components as the only covariates. Typical Cox regression model with all factors and all pair-wise interactions did not converge due to non-estimability. For cross validation performance, the 8 PLS components had to be used as the only covariates for the Cox PH model; the coefficients obtained from the PLS Cox model were set as the initial coefficients corresponding to the 8 PLS components for the Cox PH model. During the cross validation, the coefficients of the 8 PLS components were derived from the Cox PH model instead of from the PLS Cox model, therefore the results may not fully represent the actual performance of the PLS Cox model.

Table 43. Model Performance of PLS Cox Regression Model with Linear Terms (1st Degree Polynomials) – Simulation Study

	Index.Orig	Train	Test	Optim	Index.Corrected
Dxy	-0.212	-0.212	-0.209	-0.003	-0.209
R2	0.102	0.102	0.1008	0.002	0.101
Slope	1.000	1.000	1.006	-0.006	1.006
D	0.010	0.010	0.015	-0.005	0.015
U	0.000	0.000	0.001	-0.001	0.001
Q	0.010	0.010	0.014	-0.004	0.014
G	0.454	0.454	0.455	0.001	0.445

The cross validated PLS Cox model was then assessed against the test set; the

prediction errors and the time-dependent AUCs were obtained similar to the other approaches; results are presented in Table 44; and the prediction errors, time-dependent AUCs and the corresponding 95% PCIs are displayed in Figure 69.

Table 44. Prediction Performance of PLS Cox Model with linear forms of all variables – Simulation Study Test Set

Yrs	Pred Errors (95% PCI)	AUCs (95% PCI)
1	0.1060 (0.0867, 0.1267)	0.6369 (0.5708, 0.7003)
2	0.1634 (0.1457, 0.1823)	0.6126 (0.5613, 0.6623)
3	0.2021 (0.1883, 0.2164)	0.5896 (0.5450, 0.6324)
4	0.2289 (0.2198, 0.2387)	0.5859 (0.5506, 0.6229)
5	0.2399 (0.2309, 0.2489)	0.5827 (0.5480, 0.6179)
6	0.2421 (0.2319, 0.2529)	0.5843 (0.5498, 0.6182)
7	0.2378 (0.2238, 0.2527)	0.5887 (0.5569, 0.6203)
8	0.2343 (0.2166, 0.2517)	0.5875 (0.5563, 0.6180)
9	0.2269 (0.2057, 0.2482)	0.5876 (0.5566, 0.6176)
10	0.2170 (0.1912, 0.2414)	0.5875 (0.5572, 0.6170)
11	0.2035 (0.1730, 0.2297)	0.5891 (0.5589, 0.6174)
12	0.1965 (0.1629, 0.2262)	0.5884 (0.5583, 0.6162)
13	0.1872 (0.1494, 0.2197)	0.5883 (0.5588, 0.6160)
14	0.1807 (0.1418, 0.2185)	0.5878 (0.5586, 0.6155)

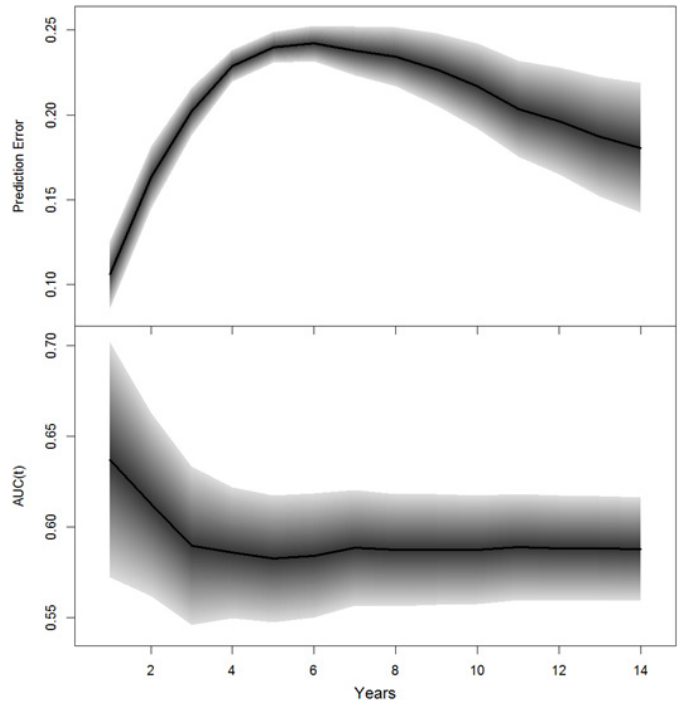


Figure 69. Prediction Performance of PLS Cox Regression with Linear Terms (1st-Degree Polynomial) – Simulation Study Test Set

4.1.2.3.7 Prediction Performance Comparison of Intended Survival Models for the Simulation Study

Thus far, the Cox PH linear model with all factors in their original scale and all potential interactions, multivariate Cox regression models with *RCS* or *FP* transformations, log-rank based RSF model, conditional-inference based RSF model, lasso, ridge, elastic-net, principal component and partial least squares Cox regression models were evaluated.

The prediction errors for all intended models are displayed again in Figure 70 for cross comparison. Of all intended models, Cox model with *FP* transformation had the best prediction error at the beginning (before year 10), but it caught up with the rest of the models at the tails, which was due to the inclusion of the time-dependent treatment interaction term; the Cox PH linear model, PCR and CINF-RSF models were the next, followed by the elastic-net Cox model and Cox model with *RCS* transformations; the lasso and PLS Cox models had almost the worst prediction errors (other than the LR-RSF model); the LR-RSF model had the worst prediction errors between 5 to 10 years, otherwise it had similar prediction errors to the CINF-RSF model.

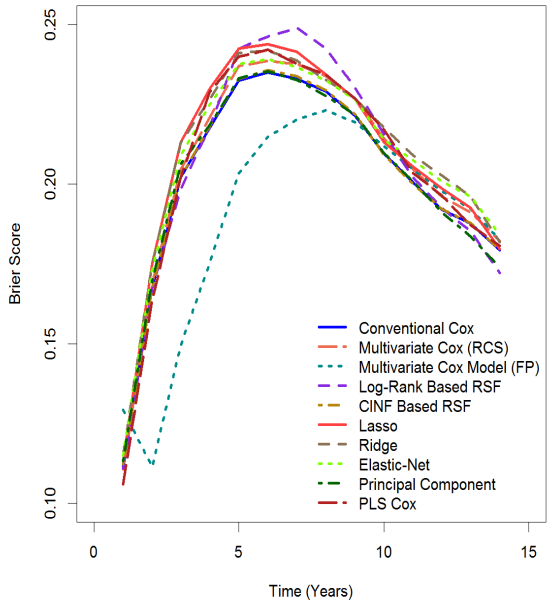


Figure 70. Comparisons on Prediction Errors – Simulation Study

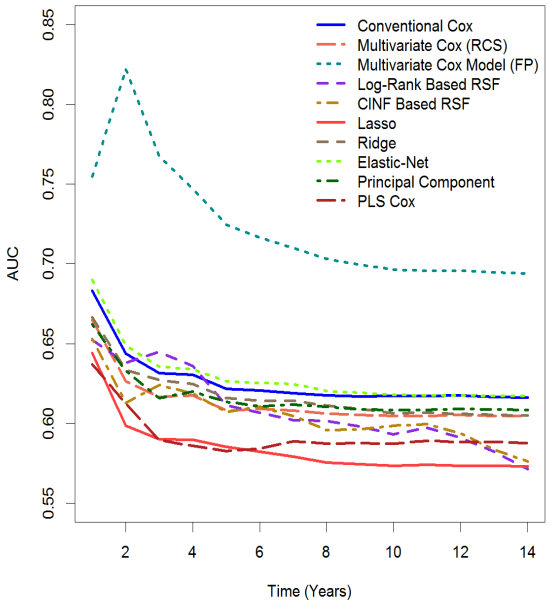


Figure 71. Comparisons on Time-Dependent AUC(t) – Simulation Study

Figure 71 displays the time-dependent AUCs for all the intended approaches; again,

the multivariate Cox regression with *FP* transformation had the best performance overall due to the inclusion of the time-dependent treatment interaction. Otherwise, the elastic-net Cox model was the second and followed by Cox PH linear model, ridge model, Cox model with *RCS* transformation and PCR; the CINF-RSF and LR-RSF model were the next. However, considering the nonparametric nature of the two RSF approaches, the convenience might overweigh the prediction performance in some cases; thus they should provide an alternative tool for analysis of survival data. Of all intended models, the lasso and the PLS Cox model had almost the worse prediction AUCs for this study.

For the 3 typical Cox PH models, the Cox model with *FP* transformation had the best performances in both prediction errors and time-dependent AUCs; and the Cox model with *RCS* transformation had the worst performance in both prediction errors and time-dependent AUCs.

LR- RSF had the worse prediction errors than the rest of the models in the middle of the curve, between 5 to 10 years; but at the tail (beyond 10-year), it became similar to the other survival models; while the CINF-RSF model had moderate prediction errors overall. In terms of time-dependent AUCs, both RSF models were moderate before year-10, and after year-10, they started to get worse until they became the worst at the year 15. In terms of selecting prognostic factors, the two RSF models picked up different factors; LR-RSF model ranked MAP as the most important factor, followed by Sex, Race and BMI; Age and Treatment were the least important factors; while the CINF-RSF model picked up Treatment then followed by MAP and Race. Apparently, the two RSF models had shown different selection patterns for prognostic factors, which was probably due to the moderate correlation between MAP and BMI. With respect to usability, both RSF models were the most convenient of all survival models due to the nonparametric nature.

For the penalized Cox models, elastic-net Cox regression models had the best performance, and selected 16 out of 345 covariate terms; the ridge Cox model was slightly worse than the elastic-net Cox model but better than the lasso Cox model and the model selected 342 out of 345 covariate terms; lasso Cox had the worst performance out of all intended survival models, but it selected 10 out of 245 covariate terms. Both Lasso and elastic-net had been very effective for prognostic factor selections.

The performance of the PCR model was also moderate, and it was only able to select

prognostic factors through loading scores for the latent components, therefore the results from this model were not intuitively interpretable.

PLS Cox regression model was an innovative approach derived for highly correlated survival data; but the prediction performance for this model was not as good as most of the intended survival models; this was possibly due to the system errors involved in the evaluation of the predicted survival probability.

Both the prediction errors and time-dependent AUCs were calculated based on the predicted survival probability estimated from a typical Cox model with the PLS components as the covariates. Instead of having the typical Cox model to solve for the coefficient estimate, the coefficients obtained from the PLS model were assigned to the corresponding PLS components as the coefficients for the Cox regression model; further, the Cox regression model was forced to take the assigned coefficients without updating them.

For evaluation of the prediction errors and time-dependent AUCs, the predicted survival probability should be based on the Cox model with coefficients corresponding to the PLS components at different time points. However, the survival status for each patient should be different at different time points, then the PLS components should be slightly different (for example, if a subject had an event at year-10, but the same subject had to be event-free at year-1; as such, the PLS components should be different at year-1 from those constructed at year-10). However, the Cox PH model itself was unable to automatically update the PLS components accordingly, it was only able to predict the survival based on the pre-assigned PLS components from the PLS Cox model. Therefore the prediction errors and time-dependent AUCs calculated from the predicted survival for the PLS Cox model were not as good as the other survival models.

In addition, another problem of PLS model was noticed in the analysis. In some extreme cases, the PLS Cox model might not be able to construct the PLS components unless all covariates were normalized; since normalized covariates should help to achieve convergence efficiently. But normalization of all covariates could potentially inflate the noise (covariates) within the dataset, thus the PLS Cox model could have picked up more noise variable than any other approaches that did not need normalized covariates; as a result, the PLS Cox model with normalized covariates tend to overfit the data.

4.2 Real World Case Study

4.2.1 NKI70 Data from Netherlands Breast Cancer Institute

A breast cancer data with 5 clinical factors and 70 gene signatures downloaded from the Netherlands Cancer Institute was used to evaluate metastasis-free survival, which will be called NKI70[143] data onward. The data included a total of 144 independent lymph-node-positive breast cancer subjects, followed for 17 months; of the 144 subjects, 48 subjects experienced metastasis since the start of the study.

This type of data reflects a new trend in today's clinical and statistical practice. In the recent years, the advances in technology and genetic analysis tools have enabled collection of large amount of genetic information; as a result, the total number of genes can be much more than the number of observations or the number of events for survival data. Such data has become very typical in biomedical research or genetic lab, especially for microarray analysis. A recent example was the study published by Beer et al. in 2002, which was performed on lung adenocarcinoma microarray expression data, the expression data were collected from 86 subjects with 7129 probe sets[153]. Another similar study published by Bhattacharjee et al. in 2001, was designed to use mRNA expressions to reveal distinct adenocarcinoma subclasses. A third study published by Garber et al. in 2001, was to study the diversity of lung adenocarcinoma with gene expressions. Many other similar studies can be found in literature. However, typical statistical models do not work for such scenarios; thus efficient analysis tools for correlated high dimensional survival data have become very demanding.

The NKI70 data were chosen intentionally for this case study so that the performance of different survival models could be studied for survival data when the number factors are more than or close to the total number of event available ($p \gg N$).

4.2.1.1 Summary Statistics of the NKI70 Data

Considering it was impossible to present all factors in one page and it was probably overwhelmed to look at the descriptive summary for all 75 factors; only the 5 clinical factors as well as 2 of the 70 gene signatures are summarized in Table 45. Contingency tables for categorical factors are presented in the top half of the table. Continuous factors,

including Age and 2 of the 70 gene signatures are summarized in the bottom half. Complete summary of all gene signatures can be found in Appendix 6.

Table 45. Brief Summary of the NKI70 Data

Factors					**Statistics (N=144)**
Survival: Metastasis					48 (33%)
Diameter of Tumor					
≤ 2 cm					73 (51%)
> 2 cm					71 (49%)
Number of affected lymph nodes					
1-3					106 (74%)
≥ 4					38 (26%)
Estrogen receptor status					
Negative					27 (19%)
Positive					115 (80%)
Missing					2 (1%)
Grade of the tumor					
Poorly Differentiated					48 (33%)
Intermediate					52 (38%)
Well Differentiated					41 (28%)
Missing					3 (2%)
Factors (71)	n/nmiss	Mean ± SD	Median	Quartiles	Ranges
Age	142/2	44.31 ± 5.34	45	41, 49	16 – 53
TSPYL5	144/0	-0.109 ± 0.33	-0.089	-0.331, 0.117	-1.08 – 0.6018
DIAPH3	144/0	-0.033 ± 0.24	-0.022	-0.179, 0.241	-0.679, 0.618
⋮	⋮	⋮	⋮	⋮	⋮
C20OR46	144/0	-0.086 ± 0.25	-0.133	-0.256, 0.020	-0.451 – 0.992

For all 75 factors, it was impossible to perform visual or manual check of the model assumptions for the Cox regression model; moreover, typical Cox models were incapable of handling so many predictors with such small number of events, thus typical Cox regression models could not be applied. For this case study, only the two RSF models, three penalized Cox regression models, PCR and PLS Cox regression models were implemented. These models were either distribution free or were capable of dealing with correlated high dimensional data; most of these models should be able to select prognostic factor, except for the ridge Cox regression, which kept most if not all factors to achieve better performance. For these approaches, there were many controversial opinions about the nonlinear transformations; in this case study, the 3 penalized Cox models and PLS Cox models were evaluated using the following 3 approaches: all factors in their original scale without interaction or transformations, all factors as well as all pair-

wise interactions and all potential 3-degree polynomial transformations for all continuous factors as well as all possible pair-wise interactions between any pair of functional forms.

For 3-degree polynomial transformations, all 71 continuous factors (including Age and the 70 gene signatures) were transformed using 3 degree polynomial forms; pair-wise interactions between any two factors (including all polynomial forms of the factors) were also constructed. Then, there would be too many covariate terms to evaluate for the survival models; for the sake of efficiency, we used hypothesis tests to prescreen potential important covariates. If a nonlinear term was significant at 0.05 level, then the term should be included as a covariate in the initial model for selection; if any one of the 3 polynomial forms of a factor was significant, then lower order forms for the same factor should also be included. For pair-wise interactions, all nonlinear forms identified from above step were used to construct the interaction terms; the interaction terms were prescreened again at 0.05 significance level. If an interaction term was significant, then the interactions should be constructed between all possible forms (including the linear and nonlinear forms of the participating factors) as identified above for the initial survival model.

4.2.1.2 Data Preparations

For this dataset, 4 clinical factors were categorical; Age and the 70 gene signatures were continuous, a good deal of the continuous factors departed from be normality. Theoretically, the departure from normality should have been fixed using transformations prior to the analysis, however transformations of gene signatures could potentially destroy the relationships among them; thus without losing too much generality, the transformations were not intended. However, in the dataset, a total of 7 missing values were observed, 2 subjects had missing Estrogen receptor status, 3 subjects had missing Grade of the tumor and 2 subjects had missing Age. The missing values for Age were imputed using the least squares multiple regression with optimum transformations; the other 5 missing categorical factors, were imputed using recursive partition (see section 3.6 for details). Additionally, one clinical factor, Grade of tumor, had 3 category levels; 2 dummy variables were created for the relative difference between categories. Thus a total of 76 factors should be considered.

Again the first step before the actual analysis is to look at the NKI70 data; the Kaplan-Meier (KM) estimates of the survival probability and Fleming-Harrington estimates of log(-log(Survival)) /or log(cumulative hazard) were plotted; results are displayed in Figure 72; the green dotted line superimposed on top of the cumulative hazard was the hazard assuming exponential survival models; where little change was observed in the slope of cumulative hazard over time. Thus, it was reasonable to consider Cox model. Figure 73 displays the correlation map of all factors; blue color indicated a positive correlation and red color indicated a negative correlation. The intensity of the color indicated the severity of collinearity. Apparently, several of the gene-signatures were highly correlated.

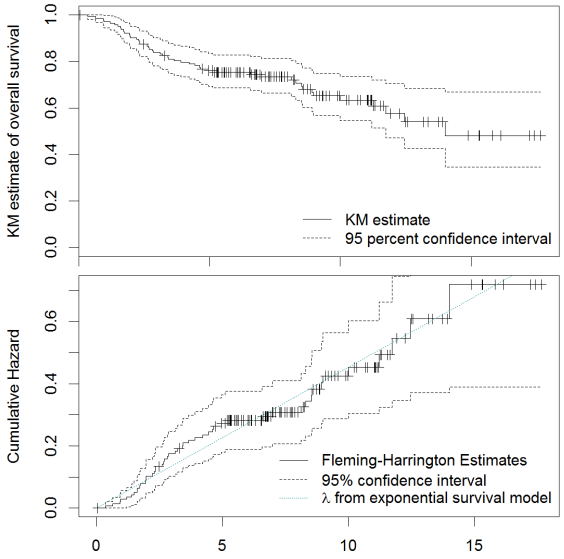


Figure 72. Kaplan Meier Survival Probability and Fleming Harrington Cumulative Hazard for NKI70 Data

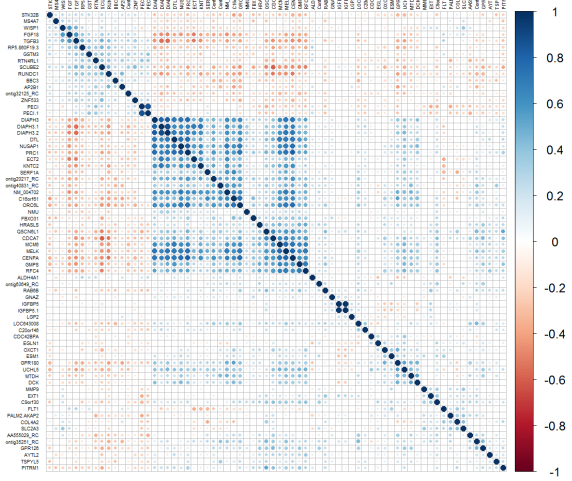


Figure 73. Factor Correlation Map for NKI70 Data

For principal component Cox regression (PCR) over all factors in their original scales, the first step was to find the minimum number of components that could achieve the majority of the total variance from all factors. Detailed summary of the component variance and the corresponding proportions are presented in Appendix 7. Figure 74 displays the total variance and the cumulative proportion of the variances contributed by each of the components; with at least 34 components, the cumulative proportion of variance can reach 90% of the total variance.

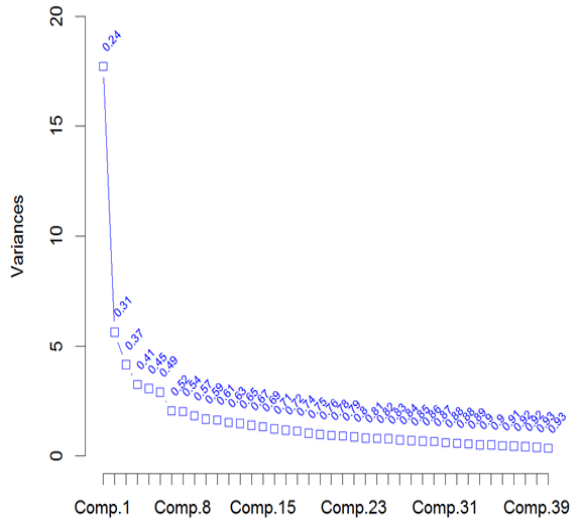


Figure 74. Total Variance against Cumulative Number of Principal Components – NKI70 Data

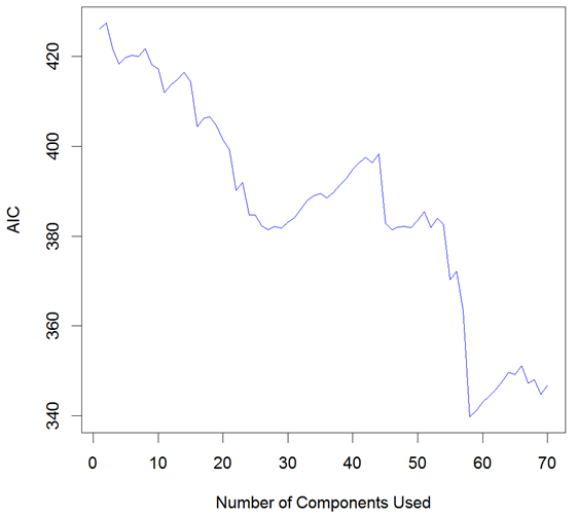


Figure 75. AIC of PCR Models against the Number of Principal Components – NKI70 Data

Figure 75 presents the AIC of the Cox regression model with the number of components included in the model; detailed summary is presented in Appendix 8. The AIC curve reached the minimum at components 58. To find out the contributions of different factors, the loadings of component 59 are presented in Figure 76; in which every 5th factor was labelled on the x-axis. The loadings had spectrum shapes, the closer to 0 in the loadings, the smaller the contributions of the factors into the component. No evident trend was noticed and the loadings of components beyond component 59 had similar patterns to what was shown in Figure 76, i.e. all of the factors had made fair contributions to the components. Looking back at Figure 74 and Figure 75, the total variance and the AIC of the Cox model were gradually decreasing with each additional component, which did not suggest exclusion of any components. Furthermore, to achieve better predictions, it was preferable to keep as many components as the model could handle. However, the more components added to the model, the less estimability. It was believed that the first 58 components were reasonable for the initial principal component Cox regression model.

Figure 77 is the hierarchical cluster analysis, which displays the linkage between factors, factors with higher correlations should probably assigned into one cluster. However, the "family" map was quite complex; it was quite trivial and time consuming to

manually construct clusters. On the other hand, the loading values of all components were used o constructed a matrix of 77 by 76 dimension; otherwise, it was not easy to get the overall picture. For similar reason, the variable clustering was skipped. Figure 77 is the hierarchical cluster analysis of all factors; highly correlated factors are linked together in the pedigree; thus the linked factors should be clustered together. However, the pedigree map was quite complex; it was quite trivial and time consuming to manually construct clusters. On the other hand, all of the intended survival models were developed for highly correlated survival data, correlated factors should be handled inside the modelling. Thus, variable clusters were not manually constructed.

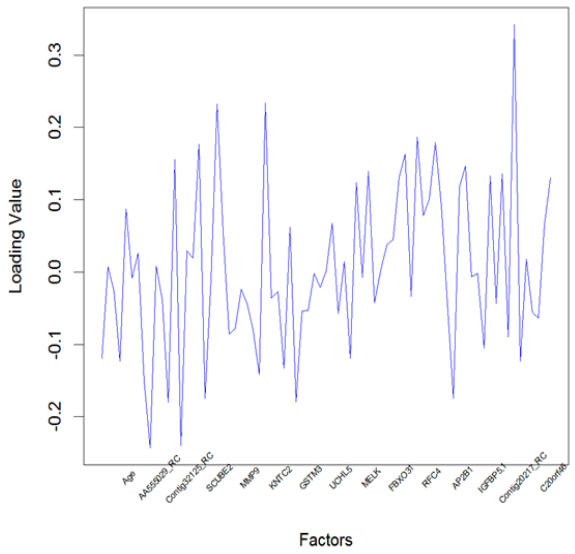


Figure 76. Loading Values of Components vs. All Factors – NKI70 Data

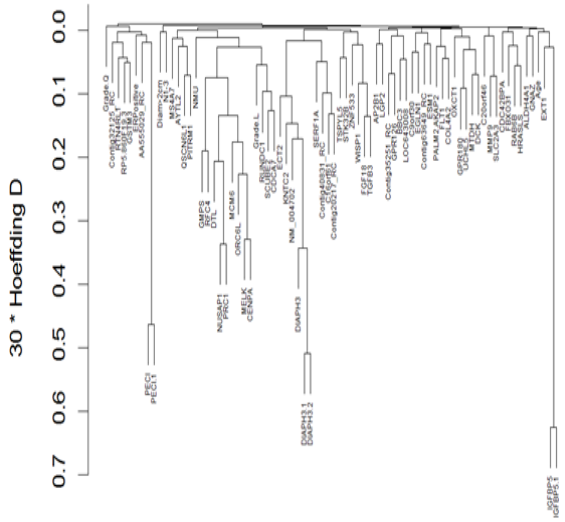


Figure 77. Hierarchical Cluster Analysis with Hoeffding's *D* statistics – NKI70 Data

Without considering nonlinearity, the linear model should include all factors in their original scale + all pair-wise interactions; thus a total of $\binom{76}{2} - 1 + 76$ =2925 covariate terms should be considered; however considering nonlinearity, the polynomial model should include all possible 3-degree polynomial forms + all pairwise interactions between the function forms of factors; thus the model should include approximately 64 billion, $[71^3 + \binom{71^3 + 5}{2} - 1 + 76]$ covariate terms. Unfortunately none of the computer software was able to handle a design matrix of dimension $(64 \times 10^9) \times (64 \times 10^9)$; thus without a proper Cox model, the residuals for every single term were not obtained. In this

case, hypothesis tests were used to prescreen potential nonlinear forms or all potential interaction terms (at 0.05 significance level). It was easy to setup hypothesis tests (see section 1.6 for the formulation of the hypotheses tests) to check for existence of polynomial forms or interaction terms. As an example:

Model 1	$S(t\|X) = {x_1}^T \beta_1$
Model 2	$S(t\|X) = {x_1}^T \beta_1 + f(x_1)^T \beta$
Model 1'	$S(t\|X) = {x_1}^T \beta_1 + {x_2}^T \beta_2$
Model 2'	$S(t\|X) = {x_1}^T \beta_1 + {x_2}^T \beta_2 + (x_1 : x_2)^T \beta_{\mathrm{I}}$

where $f(x_1)^T$ is the nonlinear functional form of x_1.

The difference in the partial log likelihood deviance between model 2 and model 1 should be the deviance corresponding to the nonlinear form of x_1, which were checked using Wald test; and the difference in the deviance between model 2' and model 1' was the deviance corresponding to the interaction terms between $x_1 : x_2$; again hypothesis was evaluated using Wald test.

For nonlinear functional forms, a quadratic form of the factors was first introduced to the Cox PH model, the deviance difference statistics between model 2 and model 1 was used to test the existence of the quadratic term; similarly, the existence of cubic form of the factors was tested using the deviance difference between the model with quadratic and cubic form of factors. For interactions terms, all functional forms of the participating factors (factors involved in the interactions) were used to construct interaction terms.

Table 46 presents the highest polynomial terms that were found to be significant from the Wald tests on the deviance difference (detailed above); the p-values from the Wald tests are also presented. A total of 19 factors were found to have significant nonlinear polynomial forms. Theoretically, if a higher order polynomial form of a particular factor was found to be significant, any lower order forms for the same factor should also be included in the initial survival model. As an example, the highest degree of the significant polynomial form for FGF18 was 3, then the cubic form, quadratic form and linear form of the gene should all be included in the initial model as well as the pair-wise interactions involving the gene signature, FGF18. Besides these predictors expressed in nonlinear polynomial forms, the rest of the 57 predictors were also included in the initial survival

model in their original scale.

Table 46. All Possible 3-Degree Polynomial Forms of Factors – NKI70 Data

Variables	Age	Contig63649_RC	QSCN6L1	FGF18	DIAPH3.1
Degrees	3	2	2	3	2
P-values	0.0448	0.0465	0.0203	0.0489	0.0478
Variables	Contig32125_RC	DIAPH3.2	RP5.860F19.3	KNTC2	WISP1
Degrees	3	2	2	2	3
P-values	0.0151	0.0163	0.0334	0.0046	0.0128
Variables	CDC42BPA	TGFB3	MELK	DTL	ORC6L
Degrees	2	2	2	2	2
P-values	0.0072	0.0187	0.0456	0.0127	0.0251
Variables	LOC643008	MCM6	PITRM1	C20orf46	
Degrees	3	2	3	2	
P-values	0.0274	0.0474	0.0464	0.0222	

For interactions, the Wald test on the deviance difference detected a total of 508 significant interactions, which were constructed from all potential polynomial forms for the 19 factor as found in Table 46, and the rest of 57 predictors in their original scale; the list of the 508 significant interactions is presented in Appendix 9. As mentioned in section 3.8.4, when an interaction between any two factors was included in the model, then all potential functional forms (as determined from the previous paragraph) from the two factors should also be included in the initial survival model and the interaction terms should be constructed by all potential functional forms of the two factors. Thus, the above 508 significant interactions, all possible polynomial forms of the 19 significant nonlinear factors and the rest of the 57 factors in their original scale formed a total of 735 covariate terms; detailed list of the 735 terms is presented in Appendix 10.

The proportionality assumptions were only tested for the 5 clinical factors; as mentioned earlier, the time-varying or time-dependent gene signatures did not make any sense, especially for the short study duration (17 months). The scaled Schoenfeld residuals plots of the 5 clinical factors were produced; the 3-level categorical factor, grade of tumor, was handled by the typical Cox model internally.

Since the 3-degree polynomial form of Age was found to be significant (see Table 46), the proportionality test for Age was carried out using linear, quadratic and cubic forms of Age; the proportionality tests for the rest of the clinical factors were performed using the factors in their original forms, since they were all categorical. The scaled

Schoenfeld residual plots for the 5 clinical factors are presented in Figure 78; the proportionality assumptions were tested using a global Chi-square test[15]; results are presented in Table 47.

Table 47. Proportionality Assumptions for All Clinical Factors – NKI70 Data

	ρ	Chisq	P-Value
Diameter	-0.3029	4.37	0.0366
N	0.2926	5.28	0.0216
ER	0.0429	0.09	0.7604
Grade	-0.2677	3.66	0.0558
Age^3	-0.4140	11.50	0.0007
GLOBAL	NA	16.62	0.0053

It can be seen from the table that proportionality tests showed significance for almost all of the clinical factors except for Estrogen Receptor Status (ER), of which the Chi-square test did not show significance due to the quadratic pattern of the scaled Schoenfeld residuals against time (shown in Figure 78), but the quadratic pattern was a clear evidence of non-proportionality. Thus, extension 2 from Section 3.8.5 was attempted for the 5 clinical factors; time-dependent interaction terms were constructed between the original factor and logarithm of time, after which proportionality assumptions were tested again, just to ensure non-proportionality had been fixed.

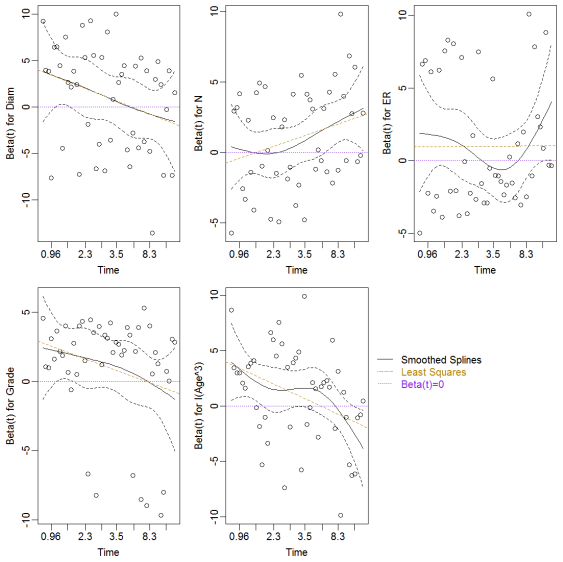


Figure 78. Scaled Schoenfeld Residuals for All Clinical Factors – NKI70 Data

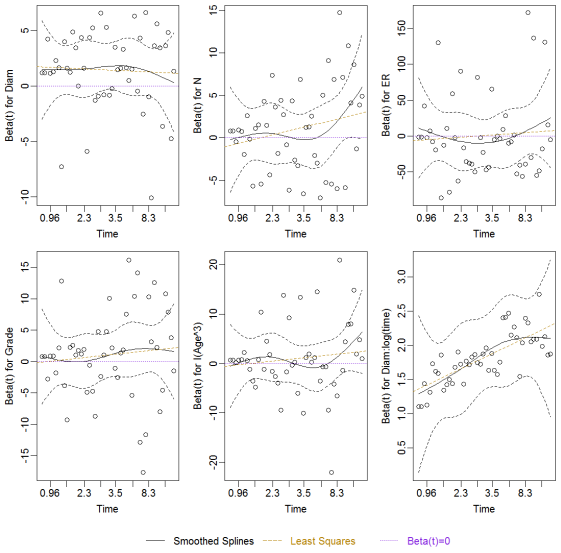


Figure 79. Scaled Schoenfeld Residuals for Clinical Factors + Diam:log(Time) from Cox Model – NKI70 Data

Table 48 presents the results of non-proportionality test for all clinical factors as well as the interaction term between the diameter of tumor and logarithm of time. Figure 79 displays the scaled Schoenfeld residuals plot for all clinical factors and the interaction term. Several factors showed slight patterns of non-proportionality, especially for the interaction between diameter of tumor:logarithm of time, which displayed some patterns of non-proportionality; but none of the proportionality tests was significant. Thus, with the addition of the time-dependent interaction term between diameter of tumor and logarithm of time, the non-proportionality issues were alleviated to allow for applications of Cox proportional hazard model.

Table 48. Proportionality Tests for Clinical Factors + Diam:log(Time) – NKI70 Data

	ρ	Chisq	P-Value
Diam	-0.049	0.07	0.7970
N	0.228	1.50	0.2200
ER	0.071	0.14	0.7040
Grade	0.095	0.34	0.5580
I(Age^3)	0.117	0.45	0.5030
Diam:log(time)	0.748	2.50	0.1140
GLOBAL	NA	4.26	0.6420

4.2.1.3 Analysis (Model Selection)

For this case study, typical Cox PH models were not able to handle so many predictors with such small number of events (due to singular design matrix). Therefore, only RSF model, penalized Cox regression (including lasso, ridge, elastic-net Cox regression), principal component (PCR) and partial least squares Cox regression models were intended. Again, the original data were randomly partitioned into training set which included a total of 32 events from 108 (75%) subjects and test set with 16 events from 36 (25%) subjects. The training set was used to train and tune the models; the test set was used for assessing predictions performance.

4.2.1.3.1 Nonparametric Random Survival Forest (RSF)

As discussed in section 3.3.1.2, Random survival forest (RSF) is a nonparametric approach with little or no assumptions, nonlinearity, non-proportionality, multicollinearity or interactions should be of no concern for RSF model; therefore it was reasonable to evaluate the two RSF models, log-rank based RSF and conditional

inference based RSF models based the training set without consideration of interactions or transformations.

4.2.1.3.1.1 Log-rank Based Random Survival Forest (RSF)

All 76 factors were fitted to the log-rank based RSF. Table 49 presents the variable importance (VIMP) of a subset of the 76 factors; one gene signature of the least importance was GPR180 with VIMP of -0.0027; 3 gene signatures with the least absolute value of VIMP were Grade.Intermediate, Contig40831.RC and PECI.1 with VIMP = 0; the factors with VIMP $\geq$ 0.0027, which was the absolute value of the least VIMP (from factor GPR180), should be potentially important. As a general rule, only factors with VIMP $\geq$ the absolute value of the least VIMP should be of interest.

Table 49. VIMP from Log-Rank Based RSF – NKI70 Data

Factors	VIMP	Relative VIMP	Factors	VIMP	Relative VIMP
ZNF533	0.0236	1.0000	Diam.GT2	0.0028	0.1167
PRC1	0.0108	0.4598	DTL	0.0027	0.1165
QSCN6L1	0.0070	0.2960	⋮	⋮	⋮
RFC4	0.0066	0.2778	Grade.Intermediate	0.0000	0.0015
CDCA7	0.0046	0.1970	Contig40831.RC	0.0000	-0.0012
IGFBP5	0.0046	0.1934	PECI.1	0.0000	-0.0020
SLC2A3	0.0034	0.1434	⋮	⋮	⋮
N.GE4	0.0028	0.1178	GPR180	-0.0027	-0.1137

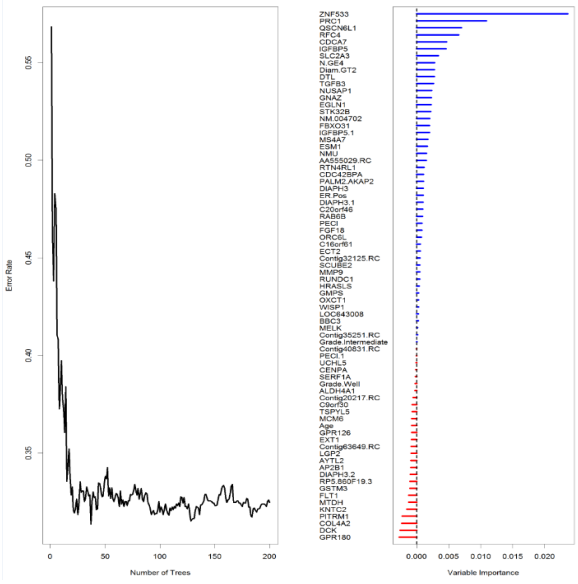


Figure 80. OOB Error Rate and VIMP from Log-Rank Based RSF – NKI70 Data

Figure 80 displays the cross-validation error and the VIMP for all factors. Based on the CV errors, the log-rank based RSF was stabilized at 25 trees and above, while the tree response still exhibited some random variations due to the nature of bootstrap aggregation. For VIMP plot, the factors were ordered from the least VIMP to the most VIMP; the factors with negative VIMP were colored in red and the ones with positive VIMP were colored in blue.

Table 50. Variable Selection from Log-Rank Based RSF – Cast Study 1

Minimum Depth Rule			Variable Hunt			Variable Importance	
Genes	Min Depth	VIMP	Genes	Min Depth	Rel Freq	Genes	Rel Freq
ZNF533	6.05	0.0236	ZNF533	2.26	100	ZNF533	98
MS4A7	6.18	0.0017	PRC1	2.72	76	PRC1	58
EXT1	6.44	-0.0008	NM.004702	3.04	52	N.GE4	56
PRC1	6.49	0.0108	IGFBP5.1	2.96	48	QSCN6L1	52
NM.004702	6.50	0.0020	IGFBP5	2.97	42	MS4A7	48
PECI.1	6.52	0.0000	QSCN6L1	2.91	40	IGFBP5	44
RTN4RL1	6.54	0.0012	PECI.1	3.36	36	GNAZ	34
IGFBP5	6.55	0.0046	CDCA7	3.32	36	ER.Pos	28
GNAZ	6.60	0.0023	EXT1	3.24	28	ECT2	26
IGFBP5.1	6.67	0.0020	CENPA	3.39	24	EGLN1	22
QSCN6L1	6.68	0.0070	OXCT1	3.23	18	ORC6L	18
AA555029.RC	6.74	0.0015	SLC2A3	3.56	18	FLT1	16
SLC2A3	6.75	0.0034	MMP9	3.43	16	CDCA7	16
WISP1	6.76	0.0003	PECI	3.25	16	DTL	12
COL4A2	6.78	-0.0023	N.GE4	5.01	14	RFC4	10
CDCA7	6.78	0.0046	Contig40831.RC	3.50	14	GMPS	8
AYTL2	6.83	-0.0009	GPR126	3.40	14	NM.004702	8
DTL	6.85	0.0027	GNAZ	3.41	12	PECI	6
ECT2	6.87	0.0005	RUNDC1	3.36	12	COL4A2	6
SCUBE2	6.90	0.0005	ER.Pos	5.77	8	EXT1	4
EGLN1	6.90	0.0023	UCHL5	3.51	8	OXCT1	4
C16orf61	6.94	0.0006	FLT1	3.58	6	BBC3	2
UCHL5	6.94	-0.0001	TGFB3	2.96	6	RP5.860F19.3	2
MMP9	6.95	0.0005	ORC6L	3.70	6	C9orf30	2
RFC4	6.95	0.0066	MS4A7	2.41	6	ESM1	2
OXCT1	6.96	0.0003	DIAPH3	4.03	2		
HRASLS	6.96	0.0004	AP2B1	3.52	2		
PECI	6.96	0.0008	EGLN1	2.71	2		

For log-rank based RSF model, important factors were selected using one of three different options, minimum depth, variable hunting and variable importance. There were slight differences amongst the results obtained with the 3 options. Table 50 presents all

factors selected with the three options. The same background color indicates the factors were consistent across all three options; the shaded background indicates the factors were consistent across two of the options. As can be seen, VIMP option selected the most factors in common with the other two options.

Additionally, Log-Rank Based RSF was able to systematically detect important pair-wise interactions, which could be useful to identify potential interactions for Cox regression analysis. Two options were available for identify potential interaction terms, maximum subtree and VIMP. However, for this cases study, a total of 2925 pair-wise interactions were constructed from the 76 factors (see section 4.2.1.2 for details) without considering the polynomial transformations, the minimum depth matrix for the maximum subtree analysis should have dimensions of 2925 × 2925, which was almost impractical to review; thus only the VIMP of the interaction was checked and the maximum subtree analysis was not performed. An example of the pair-wise interactions with relatively large VIMP as compared to either of the factors involved in the interaction is presented in Table 51.

Table 51. A Subset of All Pair-Wise Interactions with Highest VIMP from Log-rank Based – NKI70 Data

Interactions	Factor 1 ($\times 10^{-3}$)	Factor 2 ($\times 10^{-3}$)	Paired ($\times 10^{-3}$)	Additive ($\times 10^{-3}$)	Diff ($\times 10^{-3}$)	Relative Diff
COL4A2:Age	-0.1	1.8	2.9	1.7	1.1	11.00
AYTL2:SLC2A3	-0.6	-0.1	0.3	-0.7	1	10.00
QSCN6L1:SLC2A3	10.6	-0.1	9.6	10.6	-0.9	9.00
RFC4:C9orf30	8.1	-0.1	7	7.9	-0.9	9.00
HRASLS:KNTC2	0.1	-0.2	0.9	0	0.9	9.00
CENPA:ESM1	0.1	0.4	-0.1	0.6	-0.7	7.00
HRASLS:DTL	0.1	0.7	1.5	0.8	0.7	7.00
HRASLS:CDC42BPA	0.1	0	-0.6	0.1	-0.7	7.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PALM2.AKAP2:FLT1	0.4	-2.1	-1.5	-1.7	0.2	0.50
C9orf30:Contig35251.RC	-0.6	-1.5	-1.8	-2.1	0.3	0.50
⋮	⋮	⋮	⋮	⋮	⋮	⋮

During the cross validation of the log-rank based RSF model, the OOB outcomes (from the left out sample) were predicted (based on the training set). The predicted OOB survival probability, OOB cumulative hazard and OOB hazard function for a subset of 3 subjects from the training set are presented in Figure 81 and the predicted OOB survival,

OOB Brier scores and OOB mortality for all subjects from the training set are presented in Figure 82.

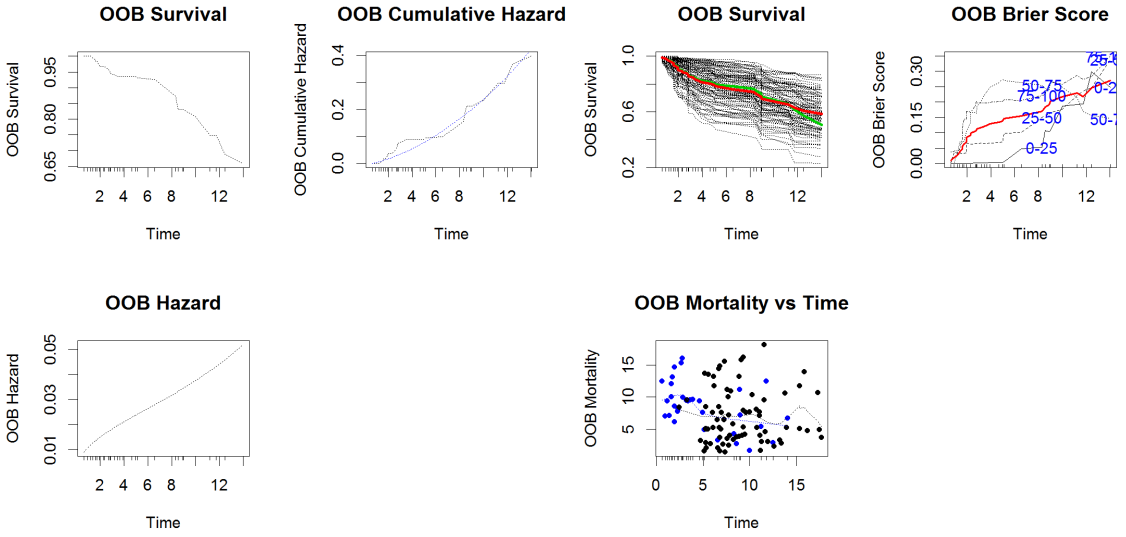


Figure 81. CV Survival, Cumulative Hazard and Hazard for LR-RSF (A Subset of 3 Subjects) – NKI70 Data

Figure 82. CV Survival, CV Error and Mortality for Log-rank Based RSF (All Subjects) – NKI70 Data

Similar to the predictions from the cross validation step, predictions on the test set were obtained from the cross validated LR-RSF model; the predicted mortality rate vs. each factor for the test set is presented in Figure 83 and the predicted survival probability against each factor for the test set is presented in Figure 84.

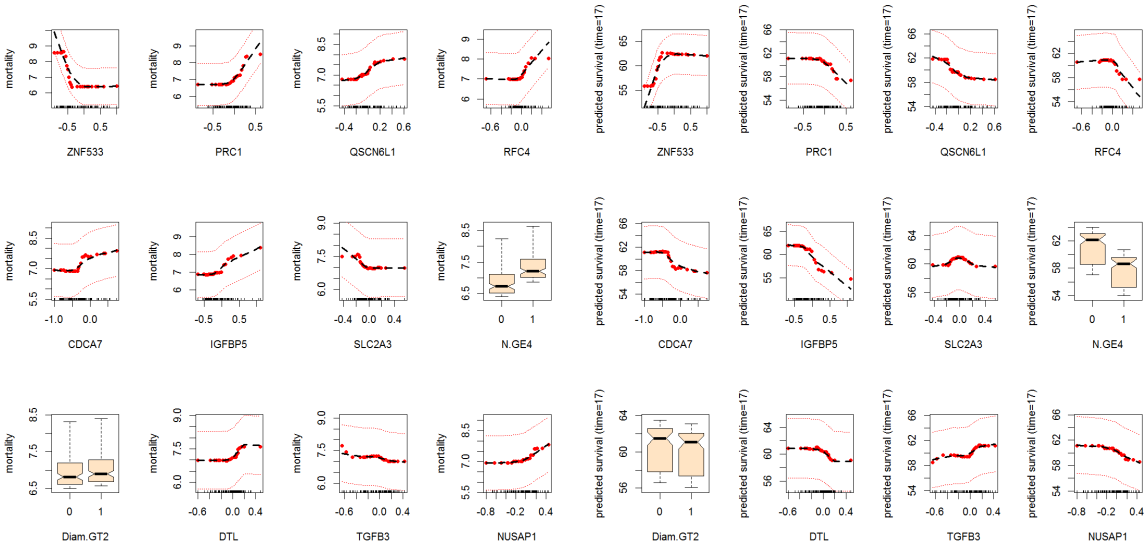


Figure 83. Predicted Mortality Rate for the first 12 Important Factors from LR-RSF – NKI70 Data

Figure 84. Predicted Survival for the first 12 Important Factor from LR-RSF – NKI70 Data

4.2.1.3.1.2 Conditional Inference Based Random Survival Forest (RSF)

The conditional inference based random survival forest (CINF-RSF) model was applied to the data with 76 factors; a conditional inference forest tree is presented in Figure 85. The CINIF-RSF model was a reasonable approach for highly correlated survival data. However, it is awkward to see negative predicted survival probably in the terminal leaves (Figure 85). In the nonparametric CINF-RSF analysis, survival probability is a step function of time; whenever there is an event, the survival probability is declined one step. For the forest tree obtained from CINF-RSF, the survival probabilities in the terminal node were estimated at the last time point; but prior to the terminal leaf, if all subjects within the same branch had already had an event, i.e., the actual survival probability might have already reached 0 before the terminals. But the predicted survival probability from the branch above the terminal leaf might not have reached 0 yet (due to incorrect predictions); thus failure events could still be predicted at the terminal leaves from the model. In other words, the cross validated CINF-RSF model could have incorrectly predicted event-free for a subject within a leaf when a metastasis event actually occurred for the same subject; therefore at the terminal leaf, incorrect predictions could have led to negative the survival probability.

In summary, the cross validated CINF-RSF model picked ZINF533, EGLN1, UCHL5, PRC1, MTDH, C20orf46, LGP2 and RTN4R1 as important prognostic factors.

The predicted survival probabilities were obtained from the cross validated conditional inference based RSF model based on the test set; the predicted survival probability and the KM curve for the test set are presented in Figure 86. There was a significant departure in the middle of the predicted survival probability curves from the Kaplan-Meier curve; the predicted survival probability curve was overly optimistic, which indirectly confirmed the cause of the negative survival probabilities in the terminal leaves. The real reason for this optimism was not completely known. It may be partly due to the extra loss of information during the split of the forest tree, since the CINF-RSF model developed forest trees based on the cutoff intervals of the predictors from the cross validation, which could have led to significant information loss due to the nature of categorizations.

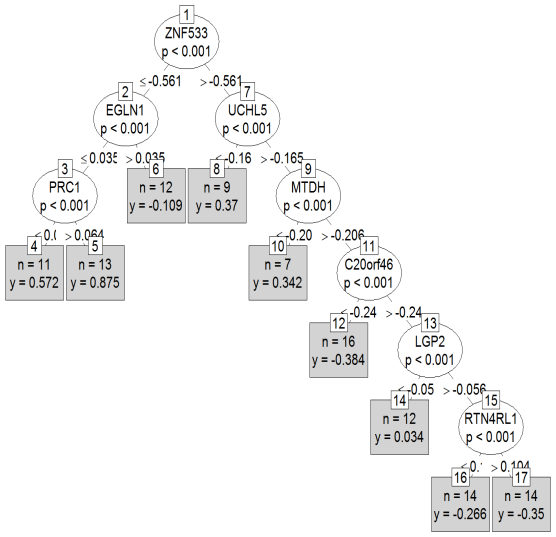


Figure 85. Forest Tree from CINF-RSF – NKI70 Data

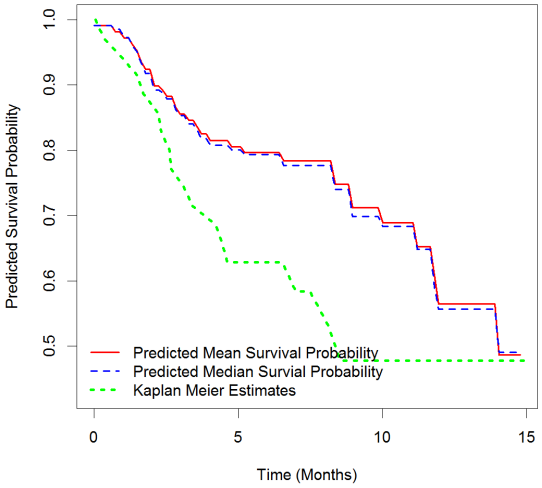


Figure 86. Prediced Survival Probability from CINF-RSF Along with Kaplan Meier Curve – NKI70 Data

The prediction performance of the cross validated models was assessed again based on the test set. The prediction errors, time-dependent AUCs of the two RSF models and the corresponding 95% PCIs were estimated based on 1000 bootstrap samples of the test set; results are presented in Table 52.

Table 52. Prediction Error and Time-Dependent AUCs for LR-RSF and CINF-RSF Models – NKI70 Data Test Set

	Prediction Errors		Time-Dependent AUCs	
Months	LR-RSF (95% PCI)	CINF-RSF (95% PCI)	LR-RSF (95% PCI)	CINF-RSF (95% PCI)
1	0.054 (0.001, 0.134)	0.052 (0.001, 0.129)	0.713 (0.553, 1.000)	0.765 (0.615, 1.000)
2	0.095 (0.031, 0.177)	0.092 (0.029, 0.172)	0.705 (0.567, 0.826)	0.756 (0.624, 0.870)
3	0.163 (0.080, 0.245)	0.155 (0.073, 0.239)	0.703 (0.547, 0.828)	0.756 (0.624, 0.868)
4	0.179 (0.109, 0.255)	0.174 (0.101, 0.251)	0.697 (0.500, 0.829)	0.757 (0.627, 0.867)
5	0.206 (0.130, 0.276)	0.213 (0.133, 0.293)	0.673 (0.500, 0.825)	0.754 (0.620, 0.867)
6	0.202 (0.129, 0.268)	0.208 (0.131, 0.286)	0.659 (0.500, 0.816)	0.753 (0.613, 0.871)
7	0.217 (0.151, 0.289)	0.224 (0.150, 0.299)	0.675 (0.500, 0.816)	0.753 (0.610, 0.866)
8	0.217 (0.151, 0.289)	0.224 (0.150, 0.299)	0.681 (0.500, 0.820)	0.755 (0.613, 0.869)
9	0.224 (0.165, 0.290)	0.226 (0.161, 0.292)	0.694 (0.500, 0.820)	0.757 (0.618, 0.871)
10	0.223 (0.169, 0.283)	0.223 (0.164, 0.284)	0.694 (0.538, 0.819)	0.757 (0.618, 0.869)
11	0.223 (0.169, 0.283)	0.223 (0.164, 0.284)	0.697 (0.541, 0.821)	0.758 (0.621, 0.871)
12	0.224 (0.170, 0.403)	0.214 (0.174, 0.332)	0.701 (0.558, 0.819)	0.760 (0.624, 0.871)
13	0.276 (0.180, 0.419)	0.258 (0.178, 0.377)	0.701 (0.562, 0.829)	0.757 (0.618, 0.870)
14	0.276 (0.180, 0.419)	0.230 (0.181, 0.298)	0.701 (0.562, 0.829)	0.757 (0.618, 0.870)

Figure 87 presents the plot of the prediction errors and the corresponding 95% PCIs and Figure 88 presents the plot of the time-dependent AUCs and the corresponding 95% PCIs. For cross comparison, the performance of the cross validated LR-RSF and CINF-

RSF models are superimposed on top of each other except with different colors. In terms of prediction errors, the two RSF models were almost equivalent most of the time, but at the tail, the conditional inference based RSF model was slightly better. In terms of the time-dependent AUCs, the conditional inference based RSF model was consistently better than the log-rank based RSF model, and the difference was quite substantial.

Prediction Errors

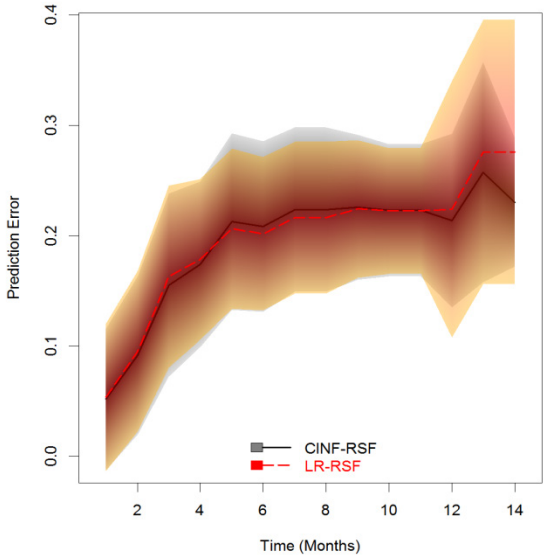


Figure 87. Prediction Error for LR-RSF and CINF-RSF – NKI70 Test Set

Time-Dependent AUCs

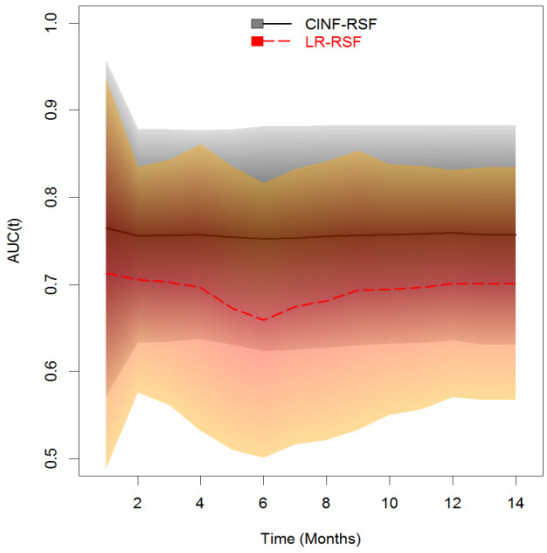


Figure 88. Time-Dependent AUC(t) for LR-RSF and CINF-RSF – NKI70 Test Set

It has to be noted that in this case study, a slight different algorithm based on asymptotic predictions was implemented for calculation of time-dependent AUCs. Since the test set only included a total of 36 subjects, of which 16 subjects had experienced metastasis and 2 of which had metastasis within the 1st months. To obtain the 95% PCIs for time-dependent AUCs, bootstrap approach was employed, where the 36 subjects were sampled with replacement. Then it was highly possible that the first two occurrences of metastasis from the test set did not make into one of the bootstrap samples; for this particular sample, the first event could occur any time after month 1; then the predicted survival time for each subject would have been infinity at month 1, the according to the definition of AUC for survival outcome (see section 3.11 for detailed definition of AUC), the AUCs could not be estimated at month 1 for this sample. Consequently, the exact

approach employed in the simulation study for estimating the time-dependent AUC would not work for this case study; therefore an asymptotic method for estimating the time-dependent AUC was developed instead: within each random sample, the scheduled AUC evaluation time was compared against the time of the next occurrence of metastasis event, the AUC(s) at the evaluation time point(s) was (/were) replaced by the AUC at occurrence of the next metastasis event, unfortunately the predicted survival probability at the occurrence of the next metastasis event should always be no better than the predicted survival probability at the evaluation time point(s). Therefore, after the replacement, the asymptotic AUCs should always be more pessimistic than the actual AUCs of the model, but considering the same method would be used to assess all intended survival models for this case study, it should still provide reasonable evidence for comparisons in this study. Though the asymptotic method was a little bit pessimistic, but with this method, the time-dependent AUCs (mean) and the corresponding 95 percentile credible intervals were obtained for each time point based on the 1000 bootstrap samples (200 bootstrap samples for RSF models).

4.2.1.3.2 Penalized Cox Regression Models

For evaluation of the penalized Cox regression models, three options were intended. The first option was to model all 76 factors or gene signatures in their original scale without any transformation or interaction terms. The second option was to model all factors in their original scale as well as all pair-wise interactions; thus total of 2925 terms (see section 4.2.1.2 for details) were included as covariates for the penalized Cox models. The last option was planned to account for all potential 3-degree polynomial transformations of the 19 factors whose nonlinear forms were found to be significant (see section 4.2.1.2 for details), the linear form for the rest of the 57 factors as well as 508 significant interactions terms, thus a total of 735 covariate terms were considered for the last option (see section 4.2.1.2). The 3 options will be referred to as linear model, interaction model and polynomial model, respectively, in this case study.

Similar to what was done in the simulation study, the penalization terms for lasso Cox regression model can obtained via CV using partial log likelihood deviance as the selection rule by setting $\alpha = 1$ and $\alpha = 0$ for ridge, but the search grid via partial log likelihood deviance is not stable, the search results are dependent on the process (results

can be slightly different based on different seeds), but the CV should be able to obtain the penalization terms when there are more factors than the number of observations available. On the other hand, the CV via Brier score is quite robust and the search paths for the penalization terms are very stable; however if the total number of observations are less than the number of factors, the search paths based on Brier scores are not able to retrieve the penalization terms for lasso and ridge Cox regression models.

In this case study, the total number of covariate terms well exceeded the total number of observations from the training set (108 subjects). Unfortunately, the cross validation via Brier score could take at most 108 factors for the lasso and ridge Cox interaction and polynomial models, thus the interaction and polynomial options for lasso and ridge models could not be cross validated via Brier score. Then CV with partial log likelihood deviance was employed instead. For all elastic-net Cox models, the cross validation via interval search was proved to be very reliable and robust from the simulation study, therefore the interval search algorithm was employed to retrieve the penalization parameters (α and λ) corresponding to the minimum cross-validation errors. For elastic-net model, there was no restriction on the total number of covariate terms to be included in the model.

4.2.1.3.2.1 Lasso Cox Models

4.2.1.3.2.1.1 Lasso Cox Linear Model

For lasso Cox linear model, a total of 76 factors (the categorical factor, Grade of tumor had 3 levels, 2 dummy variables were previously created for the relative difference within the 3 categories) were considered; the number of factors was less than the total number of observations (108), the CV via Brier score was employed. The coefficient solution paths for all factors included in the lasso Cox regression linear model are displayed in Figure 89; an increasing λ could shrink all coefficients towards 0; i.e., if λ is sufficiently large, all coefficients could be shrunk to 0, in which case the model would have no covariates left; smaller penalization term (λ) could keep more covariates. When λ was close to 0, the performance of the lasso model would be dominated by ridge regression; most (if not all) of the covariates should be kept in the model. For lasso Cox linear model, cross validation via Brier score was employed to retrieve the penalization

terms corresponding to the minimum CV errors. As a result, the penalization term, λ = 0.0441, was selected corresponding to the best performance (minimum Brier scores); the cross validation error (Brier score) curves are displayed in Figure 90.

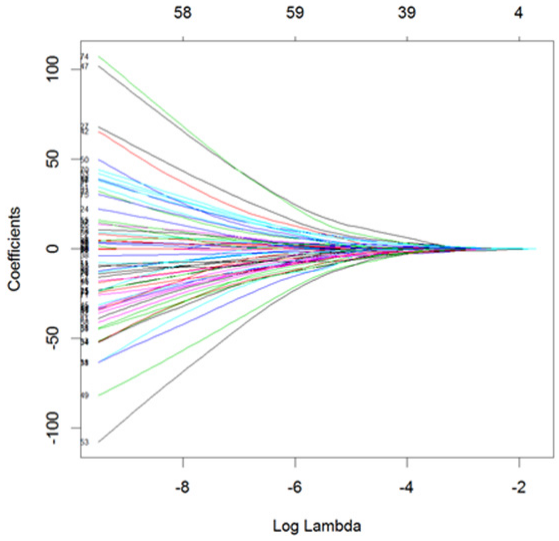


Figure 89. Coefficients Solution Path for all Factors – NKI70 Data

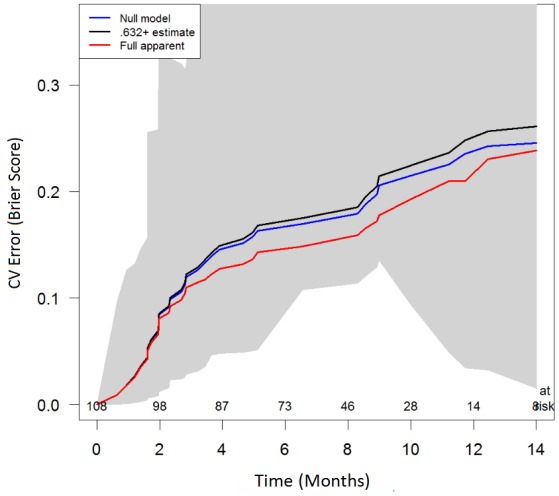


Figure 90. CV Errors for Lasso Cox Linear Model – NKI70 Data

The coefficients obtained from the cross validated lasso Cox linear model are presented in Table 53, as a comparison, the unbiased coefficients estimates for the corresponding factors obtained from a typical Cox regression model are also presented. From the table, it can be seen that smaller coefficients (in absolute values), should have less bias, the bigger the absolute values of the coefficients, the bigger the bias. For lasso Cox linear model, 3 factors were selected, Diam.GT2, Age and ZNF533; of which, Age had the smallest coefficient (0.04) in absolute values, the unbiased coefficient estimate for Age was 0.05, the observed bias was very small; for ZNF533, the coefficient estimate was -0.6451, the absolute value of the coefficient was the biggest; the unbiased coefficient estimate was -1.6468; the bias was quite substantial. The forest plot of the estimated hazard ratios from Lasso Cox linear model is presented in Figure 91.

Table 53. Biased of Coefficients from Lasso (λ=0.0441) Cox Linear Model and Unbiased Estimates from Typical Cox Regression Model – NKI70 Data

	Lasso Cox Regression (Biased)				Cox Regression (Unbiased)			
	Coef	HR	SE (Coef)	P-val	Coef	HR	SE (Coef)	P-val
Diam.GT2	0.1075	1.1135	0.3666	0.7693	0.8182	2.2664	0.3894	0.0356
Age	-0.0407	0.9601	0.0316	0.1987	-0.0585	0.9432	0.0312	0.0607
ZNF533	-0.6451	0.5246	0.4186	0.1233	-1.6468	0.1927	0.4967	0.0009

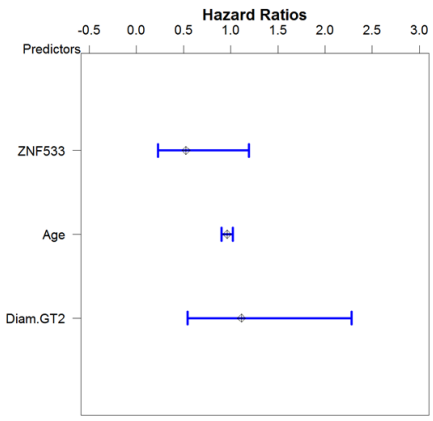


Figure 91. Forest Plot of Hazard Ratios for Lasso Linear Cox model (λ=0.0441) – NKI70 Data

The cross validated model was then evaluated against the test set to obtain the prediction performance, including prediction errors and time-dependent AUCs. For cross comparisons with the other lasso Cox models, the prediction performance will be presented later in section 4.2.1.3.2.2; the prediction errors will be presented in Table 58 and the corresponding plot will be displayed in Figure 96 and the prediction AUCs will be presented in Table 59 and the corresponding plot will be displayed in Figure 97.

4.2.1.3.2.1.2 Lasso Cox Interaction Model

Lasso Cox interaction model was intended with a total of 2925 covariate terms, including 76 factors and all pair-wise interactions; the number of covariates was much more than the total number of observations in the training set (32 events from 108 observations) or the original NKI70 dataset (48 events from 144 observations). In this case, the solution paths via Brier score did not work; instead the partial log likelihood deviance had to be used in the cross validate. The CV via partial log likelihood deviance should be able to obtain the penalization terms even if the number of covariates is more than the total number of observations, although the CV did not guarantee to achieve the global minimum of partial log likelihood deviance or global minimum of Brier scores. Unfortunately, the CV with partial log likelihood deviance was not very robust; in order to endure a better chance for finding the global minimum, the CV process was repeated 100 times, the minimum deviance and the corresponding λ and seed used within each of the 100 CV process were saved. At the end, the λ corresponding to the minimum of all saved minimum deviance from the 100 CVs should be the optimal penalization

parameter. Again, even 100 CVs still did not guarantee to find the global minimum of the partial log likelihood deviance, it should ensure a better chance to find a reasonably minimal partial log likelihood deviance.

Table 54. Top 10 Minimum Deviance from 100 CVs of Lasso Cox Interaction Models via Partial Log Deviance – NKI70 Data

Iteration	Seeds	λ	Min Deviance
70	-516583668	0.363084	9.086516
29	1524008346	0.346582	9.219817
38	-801485208	0.346582	9.230079
18	307686511	0.315792	9.263575
60	1495701791	0.346582	9.267386
25	-999248145	0.363084	9.337632
4	1954615209	0.398485	9.341521
16	-1297945748	0.437337	9.380679
90	1473660482	0.380373	9.459718
68	1977097653	0.346582	9.485587
70	-516583668	0.363084	9.086516

Table 54 only presents the results from the smallest 10 minimum deviances out of 100 CVs; Iteration is the number of cross validation in which the minimum deviance was obtained; the seeds corresponding to the 10 minimum deviances are also presented. Unfortunately, the deviance was not equivalent to the CV error; the penalization terms for the reasonably minimal partial log likelihood deviance may not correspond to the minimum of CV error.

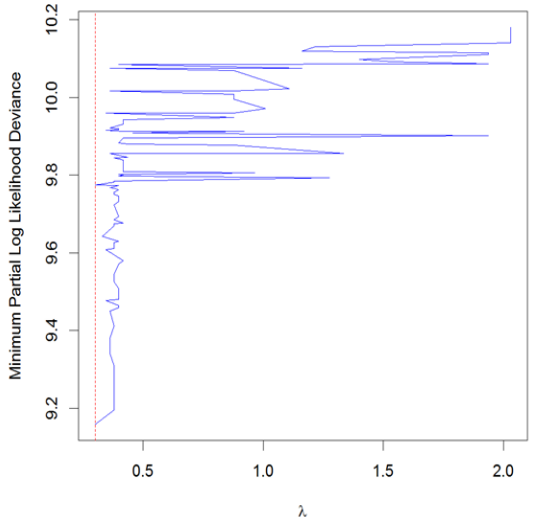


Figure 92. Minimum Deviance from 100 CVs of Lasso Cox Interaction Models vs. λ – NKI70 Data

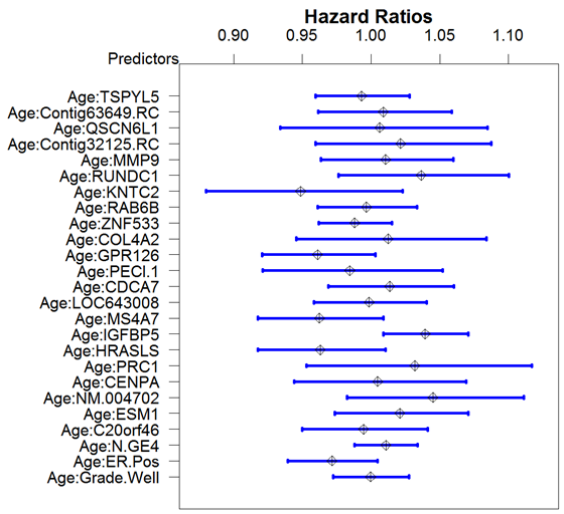


Figure 93. Biased Estimators of Hazard Ratios from Lasso Cox Interaction Model with λ = 0.3630845 – NKI70 Data

Figure 92 displays the minimum partial log likelihood deviance from the 100 CVs against the corresponding λ (presorted by the minimum deviance); the red vertical line is the λ corresponding to the reasonably minimal deviance obtained from the 100 cross validations.

As can be seen that the minimum partial log likelihood deviance was not monotonically increasing with λ; the irregular shape of the curve indicated that the deviance could have many local minimums; within each iteration, the CV process had selected a minimum partial log likelihood deviance and each minimum corresponded to a local minimum of the deviance function. Fortunately, there was a single λ valued at 0.3630845, corresponding to the reasonably minimal deviance of 9.086516, from the best lasso Cox interaction model.

With λ = 0.3630845, the lasso Cox interaction model retained 25 covariate terms; a brief summary of the coefficients corresponding to the 25 covariate terms is presented in Table 55; full summary of the coefficient estimates for all covariates from the lasso interaction model can be found in Appendix 11 and the unbiased estimates are presented in Appendix 12.

Table 55. Biased Coefficient Estimates from Lasso Cox Interaction Model with λ = 0.3631 – NKI70 Data

Terms	Coef	Terms	Coef	Terms	Coef
Age:TSPYL5	-0.0068	Age:COL4A2	0.0123	Age:PRC1	0.0313
Age:Contig63649.RC	0.0089	Age:GPR126	-0.0397	Age:CENPA	0.0047
Age:QSCN6L1	0.0064	Age:PECI.1	-0.0156	Age:NM.004702	0.0440
Age:Contig32125.RC	0.0213	Age:CDCA7	0.0135	Age:ESM1	0.0208
Age:MMP9	0.0106	Age:LOC643008	-0.0013	Age:C20orf46	-0.0055
Age:RUNDC1	0.0359	Age:MS4A7	-0.0383	Age:N.GE4	0.0107
Age:KNTC2	-0.0526	Age:IGFBP5	0.0387	Age:ER.Pos	-0.0289
Age:RAB6B	-0.0034	Age:HRASLS	-0.0377	Age:Grade.Well	-0.0003
Age:ZNF533	-0.0119				

Surprisingly, all of the 25 covariate terms as selected by the CV were interactions involving Age, which suggested that Age was probably the most important factor for the survival outcome. Forest plot of the biased estimates of the hazard ratios (exponential of the coefficients) from lasso Cox interaction model is presented in Figure 93.

The selected model was assessed against the test set to obtain the prediction performance, which will be presented in section 4.2.1.3.2.2 for cross comparison with the

other 2 lasso Cox models.

4.2.1.3.2.1.3 Lasso Cox Polynomial Model

The lasso Cox polynomial model, was intended to model the 735 covariate terms including all potential 3-degree polynomial forms of the 19 factors, the 57 factors in their original scale and all potential interactions; again, the number of covariate terms were much more than the total number of observations in the training set (32 events from 108 observations). Therefore, Brier score could not be used to cross validate the model. This model was cross validated via partial log likelihood deviance as well. Similarly, the cross validation was repeated 100 times, each with different seed. Table 56 only presents the smallest 10 minimum deviances of the 100 CVs; Iteration is the number of CV in which the minimum deviance was obtained based on partial likelihood deviance; the seeds corresponding to the minimum deviances were also saved.

Table 56. Top 10 Minimum Deviance from 100 CVs of Lasso Cox Polynomial Regression Models via Partial Log Deviance – NKI70 Data

Iteration	Seeds	λ	Min Deviance
64	-1024794212	445.440228	9.6388
93	-1693777294	709.266811	9.9326
87	-1746691985	937.608983	9.9534
3	1356247314	646.257502	9.9570
54	432562754	1638.49833	9.9592
85	2131291292	1716.51666	9.9674
92	141228043	2377.18052	9.9761
34	-1946920574	2377.18052	9.9820
62	-11704700	854.314385	9.9991
59	1904569528	2377.18052	10.0255
96	-1003577209	2377.18052	10.0288

Figure 94 displays the minimum deviance from the 100 CVs against the corresponding penalization term λ; the red vertical line is the λ corresponding to the reasonably minimal deviance of all minimum deviance obtained from the 100 cross validations. The minimum deviance curve was not necessarily increasing with increasing λ's again; however, of all penalization terms (λ's), there was a single λ valued at 445.440228 corresponding to the reasonably minimal deviance at 9.6388.

With the penalization term, λ=445.440228, the model retained three factors, Age^3, Age^3:GNAZ and Age^3:Contig40831.RC. The coefficient estimates for the 3 factors are

presented in Table 57; as a comparison, the unbiased estimates for the three coefficients estimated from a typical Cox model are also presented.

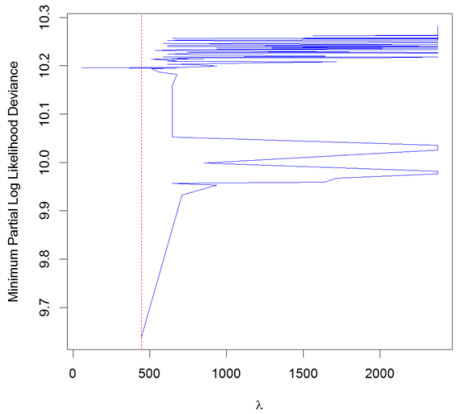


Figure 94. Minimum Deviance from 100 CVs of Lasso Cox Polynomial Regression Models vs. λ – NKI70 Data

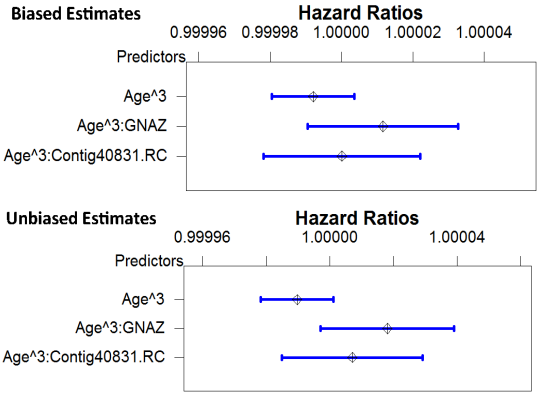


Figure 95. Forest Plot of Hazard Ratios from Lasso Cox Interaction Model with λ = 445.44 – NKI70 Data

The coefficient estimates from the lasso Cox polynomial regression were not too much biased from the ones obtained from typical Cox regression model, but substantial differences were observed in the p-values. The biased estimates of the hazard ratios from lasso Cox polynomial model are displayed in the top half of Figure 95; the bottom half of the figure presents the unbiased estimates of the hazard ratios from a typical Cox regression model. Some bias was observed, even though not very much; the 95% confidence intervals of the unbiased estimates of the hazard ratios obtained from the typical Cox regression model were slightly narrower than the ones from the lasso Cox polynomial model.

Table 57. Biased and Unbiased Coefficient Estimates from Lasso Cox Polynomial Regression Model with λ = 445.44 – NKI70 Data

	Biased Coefficients – Lasso Cox Polynomial					Unbiased Coefficients – Typical Cox Model				
	Coef	HR	SE (coef)	z	P-val	Coef	HR	SE (Coef)	z	P-val
Age^3	-7.9E-06	1.0	5.9E-6	-1.35	0.178	-1.0E-05	1.0	5.8E-6	-1.75	0.080
Age^3:GNAZ	1.2E-05	1.0	1.1E-5	1.08	0.280	1.8E-05	1.0	1.1E-5	1.69	0.091
Age^3:Contig40831.RC	1.2E-07	1.0	1.1E-5	0.01	0.991	7.1E-06	1.0	1.1E-5	0.63	0.530

4.2.1.3.2.2 Prediction Performance of Lasso Linear, Lasso Interaction and Lasso Polynomial Cox Models

The 3 lasso Cox models, lasso linear, lasso interaction and lasso polynomial with

appropriate penalization term (λ) obtained from cross validations were assessed based on the test set for prediction performance. The prediction errors and the corresponding 95% PCIs (obtained from 1000 bootstrap samples) are presented in Table 58 and the plots of prediction errors of the 3 models were superimposed on top of each other, which is displayed in Figure 96.

Table 58. Prediction Errors for Lasso Cox Models – NKI70 Test Set

Months	Lasso Linear (95% PCI)	Lasso Interaction (95% PCI)	Lasso Polynomial (95% PCI)
1	0.053 (0.000, 0.130)	0.051 (0.000, 0.128)	0.054 (0.000, 0.133)
2	0.090 (0.027, 0.175)	0.085 (0.017, 0.175)	0.095 (0.029, 0.184)
3	0.162 (0.075, 0.255)	0.145 (0.057, 0.247)	0.166 (0.077, 0.260)
4	0.187 (0.105, 0.282)	0.174 (0.084, 0.283)	0.195 (0.110, 0.290)
5	0.234 (0.148, 0.327)	0.228 (0.123, 0.336)	0.240 (0.155, 0.331)
6	0.230 (0.147, 0.319)	0.223 (0.121, 0.329)	0.236 (0.152, 0.324)
7	0.238 (0.154, 0.323)	0.238 (0.140, 0.343)	0.245 (0.161, 0.332)
8	0.238 (0.154, 0.323)	0.238 (0.140, 0.343)	0.245 (0.161, 0.332)
9	0.247 (0.182, 0.316)	0.252 (0.167, 0.348)	0.246 (0.181, 0.313)
10	0.240 (0.181, 0.302)	0.243 (0.162, 0.332)	0.238 (0.181, 0.300)
11	0.241 (0.181, 0.304)	0.243 (0.162, 0.334)	0.239 (0.181, 0.301)
12	0.234 (0.184, 0.374)	0.237 (0.150, 0.432)	0.230 (0.185, 0.367)
13	0.261 (0.194, 0.355)	0.275 (0.157, 0.437)	0.257 (0.189, 0.357)
14	0.261 (0.194, 0.355)	0.275 (0.157, 0.437)	0.257 (0.189, 0.357)

In Figure 96, the black solid curve, the red dashed curve and the blue dotted curve are the prediction errors from the lasso Cox linear, lasso Cox interaction and lasso Cox polynomial models, respectively; the gray shaded area, the orange shaded area and the greenish yellow shaded area covers the 95% PCI of the prediction errors from the lasso Cox linear, lasso Cox interaction and lasso Cox polynomial models, respectively. As can be seen, very minimal difference was observed between the three models. Specifically, within the first 9 month, the lasso Cox interaction model had the best prediction errors numerically and the lasso Cox polynomial model was the worst, however, the maximum difference of 0.02 in prediction errors between any two models was probably not much meaningful. Beyond month 9, the lasso Cox polynomial model had the best prediction errors numerically and the lasso Cox interaction model was the worst, again the maximum difference had never exceeded 0.02. In terms of the 95% PCIs, the 3 lasso Cox models almost overlapped with each other for the entire study period; whilst the lasso

Cox interaction model had the widest 95% percentile credible intervals. Overall, the prediction errors of 3 lasso cox models were almost equivalent to each other except that the lasso Cox interaction model had the worst prediction errors in the tail (beyond 12-year).

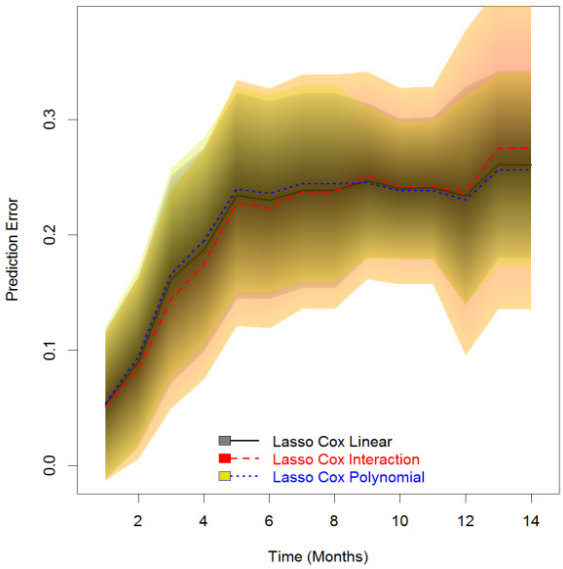


Figure 96. Prediction Errors for Lasso Cox Models – NKI70 Test Set

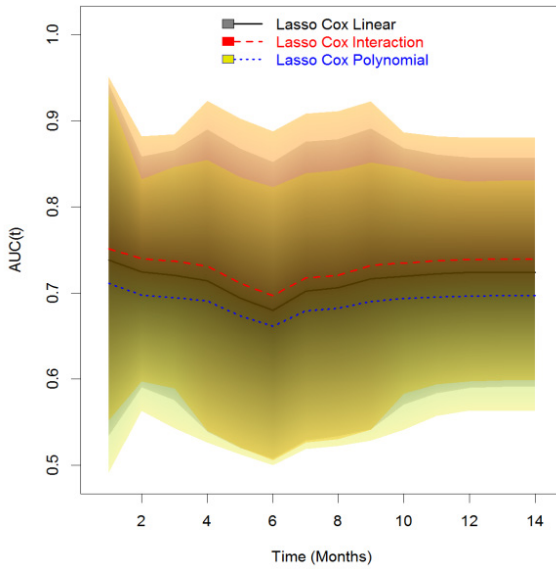


Figure 97. Time-Dependent AUCs for Lasso Cox Models– NKI70 Test Set

Besides predictor errors, model performance was also evaluated using time dependent AUCs; results are summarized in Table 59 and the plot of AUCs is displayed in Figure 97. With respect to the time-dependent AUCs, the lasso Cox polynomial model had consistently the worst performance; the lasso Cox interaction model had consistently the best performance; and the lasso Cox linear model was just in the middle. Thus, polynomial transformation of the factors did not improve the performance of lasso Cox model. Again, the difference was quite persistent among the three lasso models even though it was very minimal.

In terms of prediction errors, the 3 lasso Cox models were almost equivalent, except lasso Cox interaction model had the worst prediction errors in the tail (beyond month 12). However, in terms of time-dependent AUCs, the lasso Cox interaction model demonstrated the best performance over the other two models and the lasso Cox polynomial model had the worst performance. Therefore, polynomial transformation of input factors was probably unnecessary for lasso Cox model. On the other hand, the lasso

interaction Cox model had the best AUCs for the entire study duration, which confirmed the existence of interactions among factors.

Table 59. Prediction AUCs for Lasso Cox Models – NKI70 Test Set

Months	Lasso Linear (95% CI)	Lasso Interaction (95% CI)	Lasso Polynomial (95% CI)
1	0.738 (0.592, 1.000)	0.752 (0.602, 1.000)	0.711 (0.561, 1.000)
2	0.725 (0.587, 0.855)	0.740 (0.602, 0.886)	0.698 (0.561, 0.831)
3	0.722 (0.562, 0.852)	0.737 (0.588, 0.883)	0.695 (0.524, 0.828)
4	0.714 (0.500, 0.851)	0.732 (0.500, 0.883)	0.691 (0.500, 0.828)
5	0.694 (0.500, 0.848)	0.712 (0.500, 0.882)	0.673 (0.500, 0.822)
6	0.680 (0.500, 0.845)	0.697 (0.500, 0.882)	0.662 (0.500, 0.823)
7	0.702 (0.500, 0.847)	0.718 (0.500, 0.882)	0.679 (0.500, 0.820)
8	0.706 (0.500, 0.845)	0.721 (0.500, 0.881)	0.682 (0.500, 0.820)
9	0.716 (0.500, 0.849)	0.732 (0.500, 0.881)	0.690 (0.500, 0.823)
10	0.730 (0.554, 0.855)	0.735 (0.578, 0.882)	0.693 (0.521, 0.826)
11	0.722 (0.575, 0.852)	0.738 (0.594, 0.882)	0.696 (0.552, 0.828)
12	0.724 (0.585, 0.852)	0.739 (0.600, 0.883)	0.697 (0.560, 0.826)
13	0.724 (0.587, 0.852)	0.739 (0.601, 0.883)	0.697 (0.561, 0.828)
14	0.724 (0.587, 0.852)	0.739 (0.601, 0.883)	0.697 (0.561, 0.828)

4.2.1.3.2.3 Ridge Cox Models

4.2.1.3.2.3.1 Ridge Cox Linear Model

Ridge Cox linear model started with the original 76 factors without any transformation or interactions; the model was cross validated via Brier scores to obtain the optimal penalization term, λ. The solution paths of coefficients and penalization terms for ridge Cox linear model are not presented, but can be provided upon request. The CV errors are presented in Figure 98. The λ corresponding to the best model was 0.04111.

For this model, the CV errors were much worse than the full apparent model, which was possibly due to overfitting; the ridge Cox linear model had probably picked up too many noise covariates with the penalization term. The CV errors from ridge regression ($.632^{+}$ estimates) were estimated via the predicted survival probability from the 10th left-out sample based on the model with the penalization parameters obtained from the 9-fold CV samples. And the CV error of the "full apparent" model was obtained by fitting the full model with the penalization term estimated from the entire training set. Therefore, the CV errors for the full apparent ridge Cox linear model were different from the ones for full apparent Lasso Cox linear model, since the penalization terms were different. The

penalization parameter λ=0.04111 achieved the minimum cross validation errors for the ridge Cox linear model by keeping all covariates.

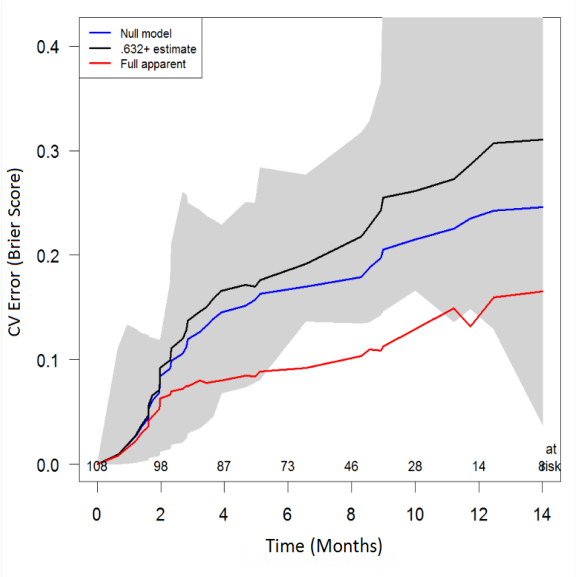


Figure 98. CV Errors for Ridge Cox Linear Regression Model with λ = 0.0411 – NKI70 Data

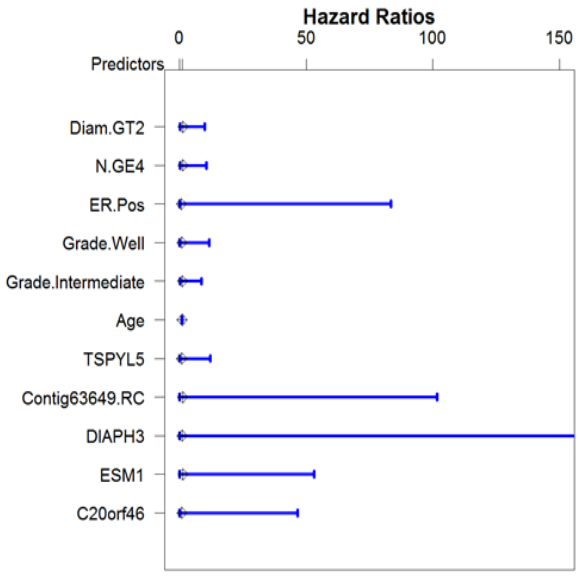


Figure 99. Forest Plot of the Subset of Hazard Ratios from Ridge Cox Linear Regression – NKI70 Data

With the penalization parameter, λ = 0.04111, the model kept all 76 factors; the coefficient estimates for a sample of 11 factors (randomly selected) are presented in Table 60; the complete summary of the coefficient estimates from the ridge Cox linear regression is presented in Appendix 11.

Table 60. A Subset of Coefficients from the Ridge Cox Linear Regression with λ =0.0411 – NKI70 Data

	Coef	HR	SE (coef)	z	P-val
Diam.GT2	0.3059	1.358	1.0199	0.2999	0.7642
N.GE4	0.2979	1.347	1.0546	0.2825	0.7776
ER.Pos	-0.3960	0.673	2.4597	-0.1610	0.8721
Grade.Well	-0.1101	0.896	1.3102	-0.0840	0.9331
Grade.Intermediate	0.0664	1.069	1.0708	0.0620	0.9505
Age	-0.0472	0.954	0.0904	-0.5219	0.6017
TSPYL5	-0.1953	0.823	1.3696	-0.1426	0.8866
Contig63649.RC	0.1990	1.220	2.2566	0.0882	0.9297
DIAPH3	0.0101	1.010	3.8526	0.0026	0.9979
⋮	⋮	⋮	⋮	⋮	⋮
ESM1	0.2628	1.301	1.8927	0.1389	0.8896
C20orf46	-0.1501	0.861	2.0366	-0.0737	0.9412

Forest plot of the hazard ratios is displayed in Figure 99; the hazard ratio for gene signature, DIAPH3, had a very wide 95% CI of [0.0005, 1921.8], which was almost 20 times wider than the second widest, therefore the confidence interval was truncated at the maximum of 150 in order to see the CIs for the rest of factors. Again, these coefficient estimates were biased. The unbiased coefficient estimates could not be obtained from the typical Cox regression model due to nonestimability; the typical Cox model could not solve for the coefficients due to singular design matrix.

The cross validated ridge Cox linear regression model was applied to the test set for assessing the predictions performance. For cross comparison with other ridge Cox models, the prediction errors will be presented later in Table 62 from section 4.2.1.3.2.3.4 and the corresponding plot will be displayed in Figure 103; the time-dependent AUCs will be presented in Table 63 and the corresponding plot will be displayed in Figure 104.

4.2.1.3.2.3.2 Ridge Cox Interaction Model

Similar to lasso Cox interaction model, the ridge Cox interaction model was cross validated via partial log likelihood deviance, which was repeated 100 times to search for the best partial log likelihood deviance and the corresponding penalization parameter (λ); the seeds and λ's corresponding to the global minimum of the minimum deviances from the 100 CVs were saved. For this model, the cross validation process was extremely unreliable (as mentioned for the simulation study), the same penalization term, λ, could result in different deviance. In this case, the same seed corresponding to the smallest deviance had to be employed in order to get the same result; fortunately, no other penalization terms could lead to the global minimum deviance. As such, the penalization term, λ, of 33.86138 with the corresponding seed of -1629866724 was able to achieve the global minimum deviance of 8.934825 (from the 100 CVs).

Figure 100 displays the solution paths for the coefficients and the penalization term (λ) for the ridge Cox interaction model. The model retained all 2925 covariate terms; all of the coefficients were either very close to zero or infinity, thus they are not be presented in the paper; a forest plot for hazard ratios for 11 random selected covariates is displayed in Figure 101.

The cross validated ridge Cox interaction model was assessed against the test set for

prediction performance; results will be presented in section 4.2.1.3.2.3.4 for cross comparison with other ridge Cox models. The prediction errors and the corresponding 95% PCIs will be summarized in Table 62 and the plot of the prediction errors will be displayed in Figure 103; the time-dependent AUCs and the corresponding 95% PCIs will be summarized in Table 63 and the corresponding plot of the AUCs will be displayed in Figure 104.

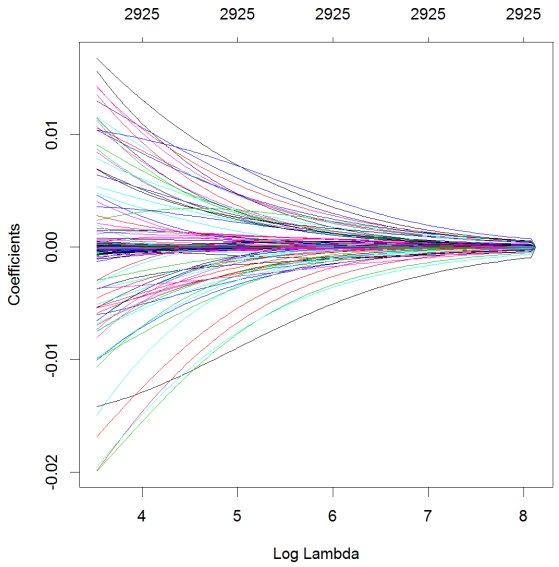


Figure 100. Solution Paths of Coefficients and λ for Ridge Cox Interaction Model – NKI70 Data

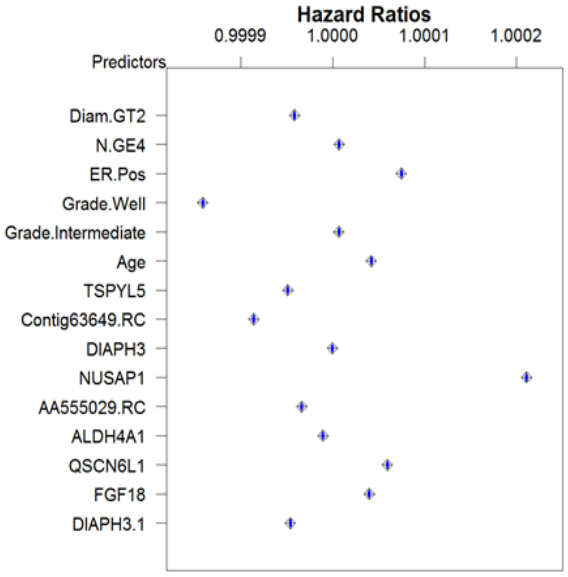


Figure 101. Forest Plot of the Subset of Hazard Ratios from Ridge Cox Interaction Model – NKI70 Data

4.2.1.3.2.3.3 Ridge Polynomial Cox Model

Similar to lasso Cox polynomial model, the CV via Brier score did not work for ridge Cox polynomial model either, because the number of covariate terms (735) were much more than the number of subjects (108) in the training set (or the original NKI70 dataset with all 144 subjects). The partial log likelihood deviance was used as the CV rule; the cross validation was repeated 100 times for searching the global minimum of the partial log likelihood deviance and the corresponding penalization term (λ).

Based on the 100 cross validations, the penalization term, λ, valued at 2377181, achieved a global minimum deviance of 9.629749 and the corresponding ridge Cox polynomial model kept all covariate terms (a total of 735 terms). Correspondingly, the coefficients were estimated from the ridge Cox polynomial model with the saved λ and

seed. The solution paths of the coefficients are presented in Figure 102.

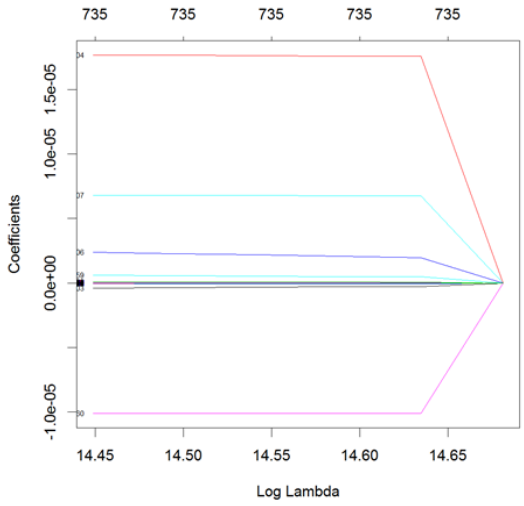


Figure 102. Solution Paths of Coefficients for Ridge Cox Polynomial Regression Model using CV via Partial Log Deviance – NKI70 Data

As mentioned in the above paragraph, the cross validated ridge Cox polynomial model kept all 735 covariate terms, only a subset of 11 random selected covariate terms are reported; and the coefficient estimates of the 11 covariate terms from the ridge Cox polynomial model are presented in Table 61; the biased estimates of the 735 coefficients from the ridge Cox polynomial model were either close to zero or infinity, the unbiased estimates however could not be obtained from typical Cox regressions due to nonestimability, therefore the complete summary of the biased or unbiased coefficient estimates are not presented. Moreover, the standard errors of the coefficients were extremely big comparing to the coefficient estimates, which should result in very wide confidence intervals, therefore the forest plot is not presented either.

Table 61. A Subset of Biased Coefficient Estimates from Ridge Cox Polynomial Model with λ = 2377181 – NKI70 Data

Factors	Coef	Factors	Coef	Factors	Coef
Diam.GT2	4.64E-38	Grade.Intermediate	5.52E-39	AA555029.RC	2.21E-39
N.GE4	5.86E-38	TSPYL5	-7.12E-40	⋮	⋮
ER.Pos	-5.10E-38	DIAPH3	1.88E-38	LGP2:C20orf46^2	7.28E-40
Grade.Well	-3.92E-38	NUSAP1	3.57E-38	CENPA:NM.004702	-1.45E-38

4.2.1.3.2.3.4 Prediction Performance of Ridge Linear, Ridge Interaction and Ridge Polynomial Cox models

The prediction performance of the three ridge Cox models, including ridge linear,

ridge interaction and ridge polynomial Cox models were evaluated using the test set, which included a total of 36 subjects and 16 events of metastasis, therefore it is expected that the 95% credible intervals should be very wide.

The prediction errors and the corresponding 95% PCIs are presented in Table 62 and the prediction errors for the 3 ridge Cox models are superimposed in Figure 103.

Table 62. Prediction Errors for Ridge Cox Models – NKI70 Test Set

Months	Ridge Linear (95% PCI)	Ridge Interaction (95% PCI)	Ridge Polynomial (95% PCI)
1	0.050 (0.000, 0.127)	0.051 (0.000, 0.129)	0.053 (0.001, 0.134)
2	0.082 (0.021, 0.161)	0.079 (0.017, 0.160)	0.094 (0.027, 0.184)
3	0.131 (0.055, 0.221)	0.131 (0.050, 0.223)	0.160 (0.073, 0.256)
4	0.152 (0.074, 0.235)	0.159 (0.079, 0.251)	0.188 (0.107, 0.274)
5	0.189 (0.110, 0.280)	0.194 (0.109, 0.294)	0.233 (0.149, 0.325)
6	0.184 (0.108, 0.273)	0.190 (0.108, 0.286)	0.230 (0.149, 0.319)
7	0.195 (0.117, 0.284)	0.202 (0.119, 0.297)	0.239 (0.162, 0.323)
8	0.195 (0.117, 0.284)	0.202 (0.119, 0.297)	0.239 (0.162, 0.323)
9	0.200 (0.134, 0.268)	0.202 (0.129, 0.280)	0.237 (0.175, 0.299)
10	0.193 (0.131, 0.254)	0.194 (0.126, 0.265)	0.229 (0.173, 0.286)
11	0.193 (0.131, 0.255)	0.194 (0.126, 0.267)	0.229 (0.173, 0.287)
12	0.192 (0.131, 0.386)	0.193 (0.123, 0.393)	0.223 (0.167, 0.382)
13	0.236 (0.141, 0.394)	0.237 (0.126, 0.403)	0.253 (0.170, 0.373)
14	0.236 (0.141, 0.394)	0.237 (0.126, 0.403)	0.253 (0.170, 0.373)

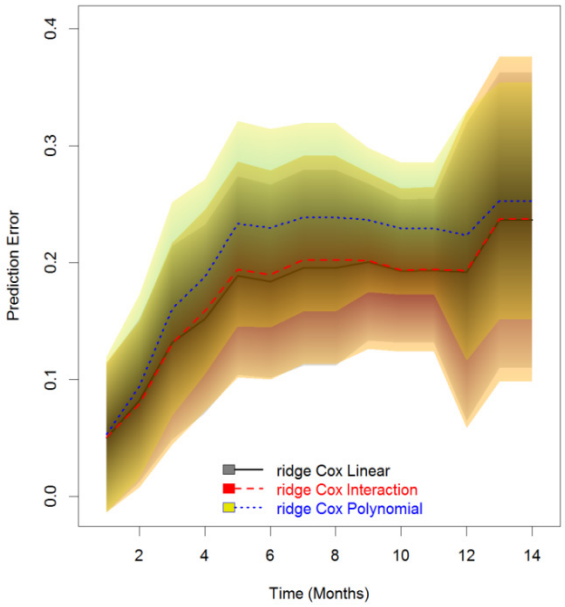


Figure 103. Prediction Errors for Ridge Cox Models – NKI70 Test Set

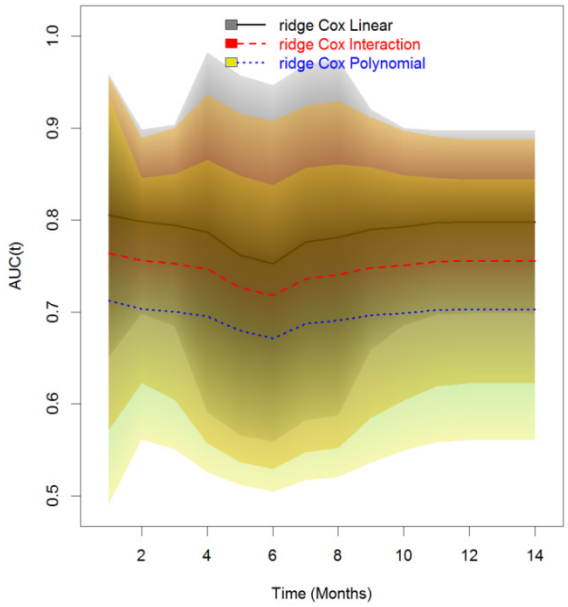


Figure 104. Prediction AUCs for Ridge Cox Models – NKI70 Test Set

In terms of prediction errors, the ridge linear and ridge interaction Cox models were

similar to each other; and both were numerically better than the ridge polynomial Cox model; though not significant, the difference was consistent over the entire study period. It suggested that polynomial transformation was unnecessary for ridge Cox models.

Besides the predictor errors, the prediction performance was evaluated using the time-dependent AUCs. The time-dependent AUCs for the 3 ridge Cox models are summarized in Table 63 and the time-dependent AUCs of the 3 models are superimposed in Figure 104. Of the 3 ridge Cox models, ridge Cox linear model had the best prediction performance with respect to the time-dependent AUCs; and the ridge polynomial Cox model had the worst prediction AUCs. Though the time-dependent AUCs among the three ridge Cox models were not significantly different, still the difference was consistent cross all time points, which suggested that the ridge Cox polynomial model and ridge Cox interaction model had probably overfitted the data.

Table 63. Prediction AUCs for Ridge Cox Models – NKI70 Test Set

Months	Ridge Linear (95% PCI)	Ridge Interaction (95% PCI)	Ridge Polynomial (95% PCI)
1	0.805 (0.693, 1.000)	0.764 (0.616, 1.000)	0.712 (0.558, 1.000)
2	0.798 (0.692, 0.893)	0.756 (0.614, 0.881)	0.704 (0.557, 0.841)
3	0.794 (0.673, 0.893)	0.752 (0.585, 0.880)	0.701 (0.541, 0.840)
4	0.787 (0.500, 0.892)	0.747 (0.500, 0.879)	0.696 (0.500, 0.840)
5	0.762 (0.500, 0.892)	0.726 (0.500, 0.879)	0.680 (0.500, 0.837)
6	0.753 (0.500, 0.888)	0.718 (0.500, 0.878)	0.671 (0.500, 0.833)
7	0.776 (0.500, 0.888)	0.736 (0.500, 0.877)	0.687 (0.500, 0.840)
8	0.781 (0.500, 0.888)	0.741 (0.500, 0.878)	0.691 (0.500, 0.840)
9	0.790 (0.627, 0.888)	0.748 (0.551, 0.878)	0.697 (0.516, 0.838)
10	0.793 (0.673, 0.888)	0.751 (0.584, 0.878)	0.699 (0.541, 0.840)
11	0.797 (0.691, 0.892)	0.755 (0.608, 0.879)	0.702 (0.553, 0.840)
12	0.798 (0.692, 0.892)	0.755 (0.614, 0.879)	0.703 (0.557, 0.840)
13	0.798 (0.692, 0.892)	0.755 (0.614, 0.879)	0.703 (0.557, 0.840)
14	0.798 (0.692, 0.892)	0.755 (0.614, 0.879)	0.703 (0.557, 0.840)

A major disadvantage of the ridge Cox regression model was that it did not perform variable selections. It kept all covariates in order to achieve "better" prediction performance; the coefficients of all covariates was shrunk towards zero with the penalization term, but never reached zero. Furthermore, it has to be noted that the prediction performances were estimated based on 16 metastasis events, therefore the 95% PCIs were extremely wide.

4.2.1.3.2.4 Elastic-Net Cox Models

Elastic-net Cox models were original build for continuous outcomes, and later it was accommodated for categorical outcomes; however little has published on survival data. In this research, the elastic-net Cox models were developed for survival outcome, three elastic net Cox models were evaluated, elastic-net linear, including the original 76 factors in the model; elastic-net Cox interaction model, include the original 76 factors as well as all pair-wise interactions (a total of 2925 covariate terms); elastic-net Cox polynomial models, including all potential 3-degree polynomial forms of the 19 factors, the linear form of the rest of the 57 factors and all potential interactions (a total of 735 covariate terms); the prediction performance was evaluated against the test set.

4.2.1.3.2.4.1 Elastic-Net Cox Linear Model

Elastic-net Cox linear model was cross validated using interval search algorithm; Figure 105 displays the last iteration of the search paths for the penalization parameters (α and λ); the model reached the best performance with the penalization parameters of α = 0.00686 and λ = 1.127488; the model retained a total of 68 covariate terms.

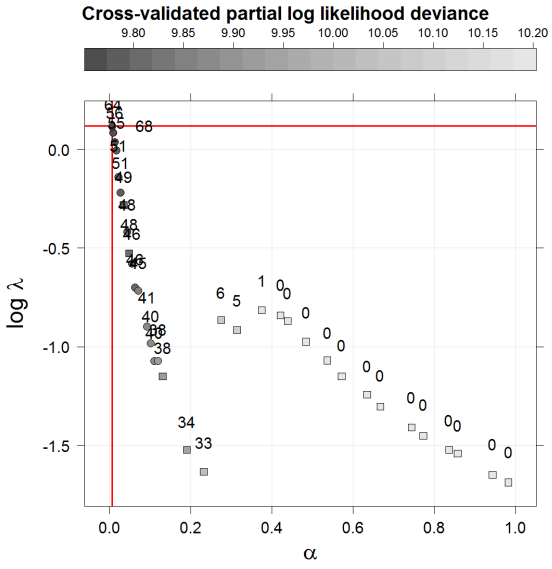


Figure 105. Interval Search Paths for Elastic Net Cox Linear Model – NKI70 Data

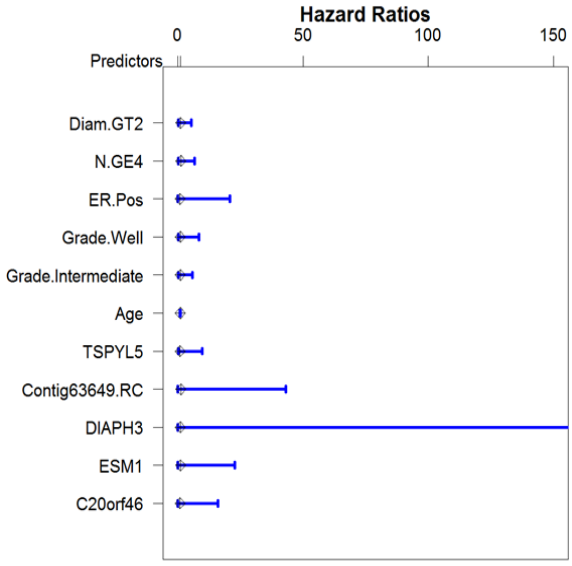


Figure 106. HR for a Subset of 11 Factors from Elastic-Net Cox Linear Model with Penalization Parameters (α = 0.0069 and λ = 1.1275) – NKI70 Data

Of all 68 covariate terms, the coefficients of 11 random selected covariate terms are

presented in Table 64; full summary of all coefficients is presented in Appendix 14. It can be seen from Table 64 that the p-values were extremely big and none were significant; since the elastic-net model selected covariates not by p-values, but by the regularization of the regression coefficients. On the other hand, the elastic-net linear Cox model selected all covariates that were highly correlated, and the number of covariates well exceeded the maximum number of parameters that could be estimated with typical Cox regressions (due to singularity); thus unbiased estimates of the coefficients could not be produced. The forest plot of hazard ratios corresponding to 11 covariate terms is displayed in Figure 106. The factor, DIAPH3, again had extremely wide confidence intervals, which was cutoff at 150.

Table 64. Subset of Coefficients for the Elastic-Net Cox Linear Model with Penalization Parameters ($\alpha = 0.00686$ and $\lambda = 1.127488$) – NKI70 Data

	Coef	HR	SE (Coef)	z	P-val
Diam.GT2	0.0840	1.088	0.8061	0.10	0.9170
N.GE4	0.1531	1.165	0.8841	0.17	0.8625
ER.Pos	-0.1239	0.883	1.6094	-0.08	0.9386
Grade.Well	-0.0468	0.954	1.1061	-0.04	0.9663
Grade.Intermediate	0.0060	1.006	0.8958	0.01	0.9946
Age	-0.0090	0.991	0.0707	-0.13	0.8983
TSPYL5	-0.0665	0.936	1.1965	-0.06	0.9557
Contig63649.RC	0.1670	1.182	1.8354	0.09	0.9275
DIAPH3	0.0144	1.015	3.0319	0.00	0.9962
⋮	⋮	⋮	⋮	⋮	⋮
ESM1	0.1222	1.130	1.5323	0.08	0.9427
C20orf46	-0.1032	0.902	1.4690	-0.07	0.9440

The cross validated elastic-net Cox linear model were checked against the test set to assess the prediction errors and the time-dependent AUCs; results will be summarized later in Table 67 and Table 68, respectively; the plots of the prediction errors and AUCs of the selected model will be presented in Figure 111 and Figure 112, respectively; for cross comparison, the elastic-net Cox linear model will be superimposed with the elastic-net Cox interaction and elastic-net Cox polynomial models in the figure.

4.2.1.3.2.4.2 Elastic-Net Cox Interaction Model

The elastic-net Cox interaction model was cross validated with interval search algorithm. Figure 107 displays the interval solution paths; surprisingly the cross

validation selected $\alpha = 0.9999$ and $\lambda = 0.1892601$, which was the very close to the penalization terms for the lasso Cox model, thus the performance of this model was dominated by the lasso Cox interaction model

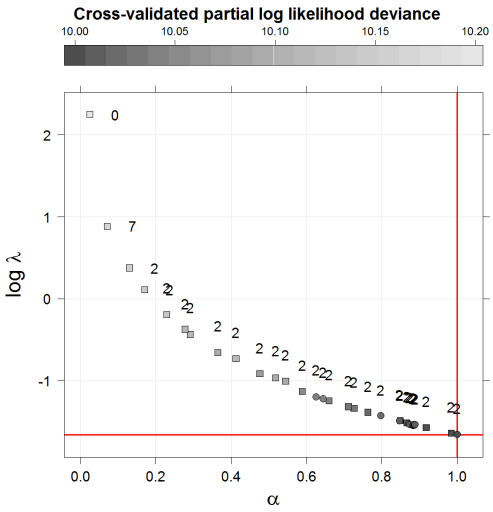


Figure 107. Solution Paths for Elastic Net Cox Interaction Model – NKI70 Data

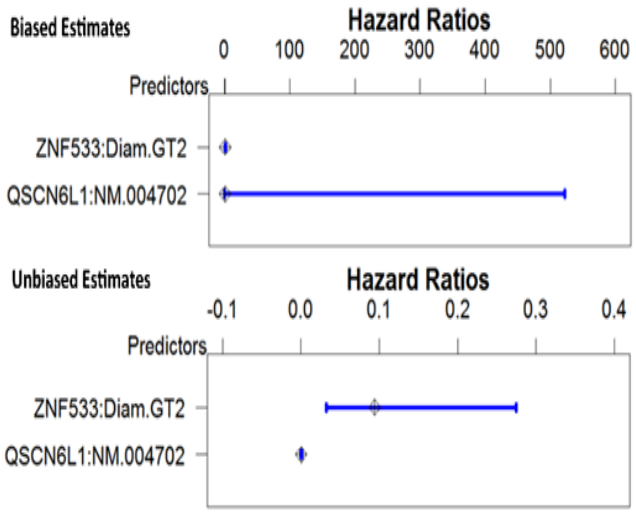


Figure 108. Hazard Ratios from the Elastic-Net Cox Interaction Model with $\alpha = 0.9999$ and $\lambda = 0.1893$ – NKI70 Data

With the obtained penalization parameters ($\alpha = 0.9999$ and $\lambda = 0.1892601$), the elastic-net Cox interaction model selected only 2 terms; the results are summarized in Table 65.

Table 65. Biased and Unbiased Coefficients for the Elastic-Net Cox Interaction Model with $\alpha = 0.9999$ and $\lambda = 0.1893$ – NKI70 Data

	Biased Estimates from Elastic-Net Cox Interaction					Unbiased Estimates from Cox Regression Model				
	Coef	HR	SE (Coef)	z	P-val	Coef	HR	SE (Coef)	z	P-val
QSCN6L1:NM.004702	-0.595	0.552	3.497	-0.17	0.865	-12.73	3.0E-6	2.900	-4.39	<.001
ZNF533:Diam.GT2	-0.513	0.598	0.547	-0.94	0.348	-2.364	0.094	0.546	-4.33	<.001

The forest plot of the hazard ratios for the selected terms is displayed in Figure 108. It can be seen from the figure that the estimates of the coefficients from the elastic-net Cox interaction were extremely biased as comparing to the ones from a typical Cox regression model. Moreover, the CV only selected 2 covariates due to the lasso-like penalization terms, which led to a lasso-like elastic-net interaction model; thus it will not be fair to compare elastic-net interaction model with the other survival models.

For cross comparison with elastic-net Cox models, the prediction performance will be reported later in section 4.2.1.3.2.4.4; prediction errors of the elastic-net Cox interaction model will be presented in Table 67 and Figure 111and the prediction AUCs will be presented in Table 68 and Figure 112.

4.2.1.3.2.4.3 Elastic-Net Cox Polynomial Model

Similarly, elastic-net polynomial model was cross validated with the training set. Figure 110 displays the interval search paths for penalization parameters; the values of the partial log likelihood deviance are presented on the top border of the figure; and the corresponding penalization parameters are presented on the axes; the number of covariates left in the model is presented next to each point. The coordinates of the intersection of the red solid lines correspond to the penalization parameters, $\alpha = 0.03704$ and $\lambda = 1.091373$, corresponding to the minimum CV errors and the model selected 226 covariate terms, which were too many to be estimated with typical Cox regressions, thus unbiased estimates of coefficients could not be obtained.

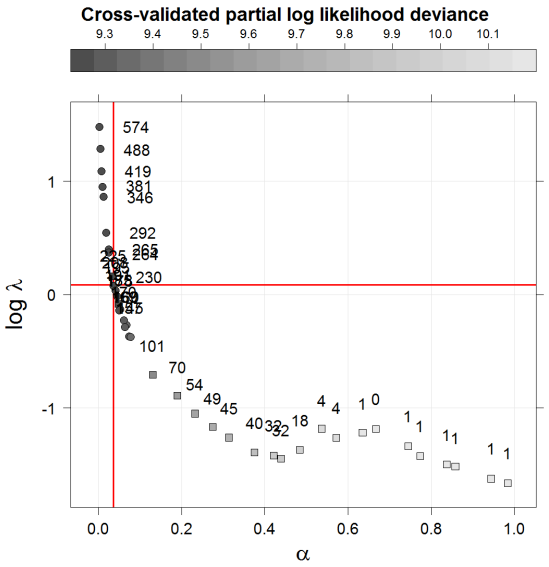


Figure 109. Solution Paths (Interval Search) for Elastic Net Cox Polynomial Model – NKI70 Data

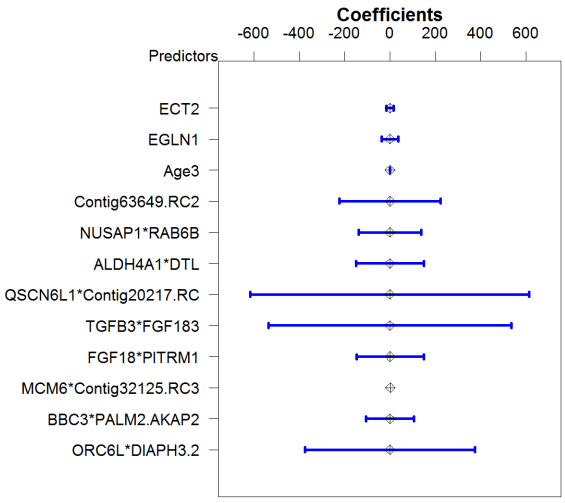


Figure 110. Forest Plot of the Subset of Coefficients from the Elastic-Net Cox Polynomial – NKI70 Data

Of the 226 covariate terms; a sample of 12 covariate terms were randomly selected and the coefficients of the 12 covariates are presented in Table 66; complete summary of

the 226 covariate terms, the corresponding coefficients, and hazard ratios are presented in Appendix 13.

Table 66. Subset of Coefficients for the Elastic-Net Polynomial Cox Model with Penalization Parameters α = 0.0370 and λ = 1.0914 – NKI70 Data

	Coef	HR	SE (Coef)	z	P-val
ECT2	0.0467	1.048	8.61	0.0054	>0.99
EGLN1	-0.1042	0.901	19.27	-0.0054	>0.99
Age^3	0.0000	1.000	0.02	0.0000	>0.99
$Contig63649.RC^2$	0.4989	1.647	113.78	0.0044	>0.99
NUSAP1:RAB6B	-0.1243	0.883	70.87	-0.0018	>0.99
ALDH4A1:DTL	-0.4106	0.663	76.47	-0.0054	>0.99
QSCN6L1:Contig20217.RC	-0.5274	0.590	314.09	-0.0017	>0.99
$TGFB3:FGF18^3$	-0.0088	0.991	273.37	0.0000	>0.99
FGF18:PITRM1	0.4207	1.523	75.94	0.0055	>0.99
$MCM6:Contig32125.RC^3$	1.3030	3.680	1.2e5	0.0001	>0.99
BBC3:PALM2.AKAP2	0.0437	1.045	53.55	0.0008	>0.99
ORC6L:DIAPH3.2	-0.1364	0.873	191.20	-0.0007	>0.99

A forest plot was produced for the coefficients (instead of the hazard ratios, since the confidence intervals for hazard ratios were too wide for this case study); the plot is displayed in Figure 110.

4.2.1.3.2.4.4 Prediction Performance of Elastic-Net Linear, Elastic-Net Interaction and Elastic-Net Polynomial Cox Models

The prediction performance of the three elastic Cox models, including elastic-net linear, elastic-net interaction and elastic-net Cox polynomial models were evaluated against the test set with a total of 36 subjects and 16 metastasis event, due to the small sample size, it is expected that the 95% PCIs were very wide for all of the 3 models.

The prediction errors and the corresponding 95% PCIs are presented in Table 67 and the plot of the time-dependent prediction errors is displayed in Figure 111. In terms of prediction errors, the elastic-net Cox polynomial model had the best performance; the elastic-net Cox linear model was the next one and the elastic-net Cox interaction model had the worst performance due to the lasso-like behavior. On the other hand, the elastic-net Cox linear and elastic-net Cox polynomial models had similar performance in terms of prediction errors, there were slight differences between the two models, the elastic-net Cox linear had slightly better prediction errors in the beginning half of the study (prior to

month 8) and had slightly worse prediction errors in the second half of the study (after month 8).

Table 67. Prediction Errors for Elastic-Net Cox Models – NKI70 Test Set

Months	E-Net Linear (95% PCI)	E-Net Interaction (95% PCI)	E-Net Polynomial (95% PCI)
1	0.045 (0.000, 0.127)	0.055 (0.000, 0.134)	0.057 (0.000, 0.139)
2	0.082 (0.021, 0.161)	0.097 (0.030, 0.186)	0.074 (0.008, 0.161)
3	0.131 (0.055, 0.221)	0.174 (0.078, 0.274)	0.137 (0.057, 0.218)
4	0.152 (0.074, 0.235)	0.201 (0.113, 0.298)	0.160 (0.085, 0.241)
5	0.189 (0.110, 0.280)	0.247 (0.162, 0.344)	0.196 (0.120, 0.277)
6	0.184 (0.108, 0.273)	0.244 (0.162, 0.337)	0.191 (0.118, 0.268)
7	0.195 (0.117, 0.284)	0.257 (0.178, 0.345)	0.202 (0.132, 0.274)
8	0.195 (0.117, 0.284)	0.257 (0.178, 0.345)	0.202 (0.132, 0.274)
9	0.200 (0.134, 0.268)	0.265 (0.200, 0.334)	0.190 (0.135, 0.249)
10	0.193 (0.131, 0.254)	0.258 (0.202, 0.321)	0.182 (0.129, 0.236)
11	0.193 (0.131, 0.255)	0.258 (0.202, 0.322)	0.182 (0.129, 0.237)
12	0.192 (0.131, 0.386)	0.251 (0.209, 0.369)	0.181 (0.127, 0.318)
13	0.236 (0.141, 0.394)	0.269 (0.219, 0.331)	0.217 (0.138, 0.330)
14	0.236 (0.141, 0.394)	0.269 (0.219, 0.331)	0.217 (0.138, 0.330)

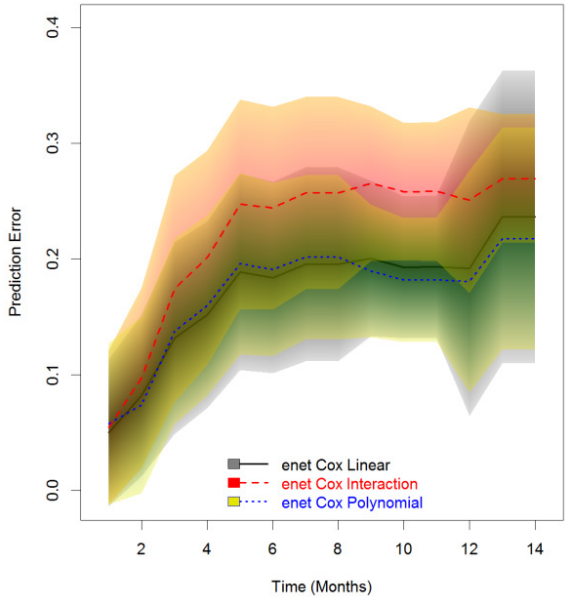


Figure 111. Prediction Errors for Elastic-Net Cox Models – NKI70 Test Set

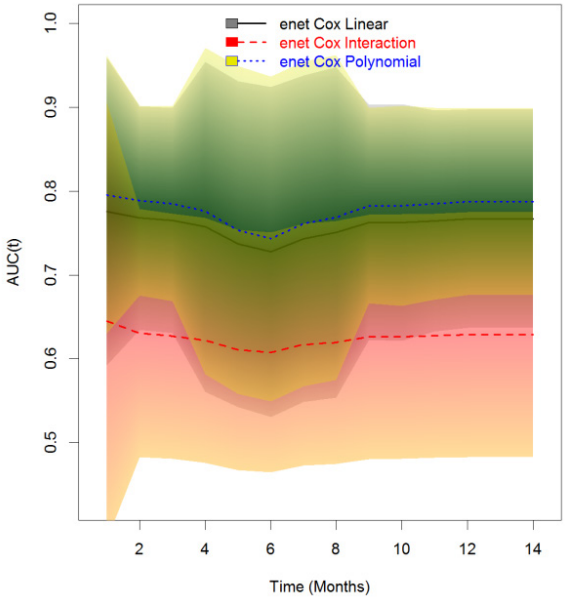


Figure 112. Prediction AUCs for Elastic-Net Cox Models – NKI70 Test Set

Besides the predictor errors, model performance was also evaluated using the time-dependent AUCs. The AUCs of the 3 elastic-net Cox models are summarized in Table 68 and the corresponding plot is displayed in Figure 112. Again, the time-dependent AUCs

of the three models were parallel for the entire study; the elastic-net Cox polynomial model had the best AUCs and the elastic-net interaction model had the worst.

Table 68. Prediction AUCs for Elastic-Net Cox Models – NKI70 Test Set

Months	E-Net Linear (95% PCI)	E-Net Interaction (95% PCI)	E-Net Polynomial (95% PCI)
1	0.776 (0.633, 1.000)	0.645 (0.477, 1.000)	0.796 (0.669, 1.000)
2	0.768 (0.633, 0.900)	0.631 (0.476, 0.771)	0.789 (0.669, 0.895)
3	0.765 (0.625, 0.894)	0.627 (0.476, 0.768)	0.785 (0.659, 0.891)
4	0.757 (0.500, 0.894)	0.622 (0.476, 0.768)	0.776 (0.500, 0.889)
5	0.737 (0.500, 0.888)	0.611 (0.478, 0.764)	0.753 (0.500, 0.891)
6	0.728 (0.500, 0.894)	0.608 (0.479, 0.764)	0.743 (0.500, 0.887)
7	0.743 (0.500, 0.889)	0.617 (0.479, 0.767)	0.761 (0.500, 0.887)
8	0.751 (0.500, 0.894)	0.620 (0.477, 0.766)	0.769 (0.500, 0.888)
9	0.763 (0.613, 0.894)	0.626 (0.476, 0.767)	0.783 (0.656, 0.889)
10	0.763 (0.611, 0.894)	0.627 (0.476, 0.767)	0.783 (0.652, 0.891)
11	0.765 (0.625, 0.889)	0.628 (0.477, 0.768)	0.785 (0.662, 0.891)
12	0.767 (0.633, 0.894)	0.629 (0.476, 0.768)	0.788 (0.669, 0.891)
13	0.767 (0.633, 0.894)	0.629 (0.476, 0.768)	0.788 (0.669, 0.891)
14	0.767 (0.633, 0.894)	0.629 (0.476, 0.768)	0.788 (0.669, 0.891)

Unlike the lasso Cox models, elastic-net polynomial Cox model had the best prediction performance; and the elastic-net interaction Cox model had the worst prediction performance. As mentioned in section 4.2.1.3.2.4.2, the cross validation for elastic-net Cox interaction model had achieved lasso-like penalization terms ($\alpha = 0.9999$ and $\lambda = 0.1892601$); which was close to those for the lasso Cox interaction model ($\alpha=1$ and $\lambda =0.3630845$). But the elastic-net Cox interaction model only retained 2 covariate terms as compared to 25 covariate terms selected for the lasso Cox interaction model, the under performance of the elastic-net Cox interaction model was probably due to underfit caused by the lasso-like penalization terms. The real reasons leading to the lasso-like penalization terms were not known, therefore it was not advisable to use elastic-net Cox interaction model for prognostic factor detection and prediction of future outcomes.

4.2.1.3.3 Principal Component Cox Regression (PCR)

For PCR, the initial step was to performance principal component analysis (PCA) to construct principal components. For the case study, a total of 76 components were constructed. As mentioned earlier in section 4.2.1.2, the original NKI70 data with all 144 subjects and 48 events was used to train the model; even so, the number of events was

still not enough to produce coefficient estimates for all 76 components, however considering that the first 70 components contributed more than 99.8% of the total variance, the first 70 components were probably enough for the initial PCR model. Since the principal components were constructed independently from the survival outcome and the total number of observations or events were barely enough to produce estimates for all 70 components, additional models to account for interaction or polynomial transformations were not implemented for the principal component Cox regression (PCR) model. Additionally, cross validation were carried out using the original NKI70 with all 144 subjects instead of the training set, then the model performance had to be evaluated based on the same data as the model was trained, thus the prediction performance was expected to be very good, but it would not be fair to compare with the other models, since the prediction performance of all the other models was assessed based on the test set.

Table 69. AIC of Deleted Components vs. df. for the Remaining PCR Model – NKI70 Data

Comp	df	AIC	Comp	df	AIC	Comp	df	AIC	Comp	df	AIC	Comp	df	AIC
Null	70	346.7	-64	65	336.7	-29	60	327.7	-66	55	321.1	-2	50	315.4
-27	69	344.7	-5	64	334.8	-21	59	326.2	-47	54	319.8	-42	49	314.9
-17	68	342.7	-70	63	332.8	-61	58	324.9	-59	53	318.7	-10	48	314.2
-63	67	340.7	-12	62	331.0	-28	57	323.6	-31	52	317.4			
-13	66	338.7	-56	61	329.3	-60	56	322.5	-40	51	316.0			

The initial PCR model included all 70 components; after backward step-down selection, 48 components remained in the model. The deleted components are presented in Table 69 in the same order as they were deleted from the model, as well as the AIC and total degrees of freedom for the corresponding model with the remaining factors.

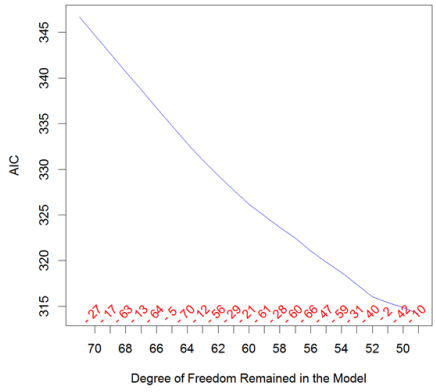


Figure 113. AIC of Deleted Components vs. df. for PCR Model – NKI70 Data

Figure 113 displays the AIC (Y-axis) of the PCR model against the corresponding degree of freedoms (X-axis); the deleted components are labelled on the X-axis in the same order as they were deleted. Detailed summary of the coefficients, hazard ratios, standard errors and the corresponding p-values for the remaining 48 components is presented in Appendix 16. Surprisingly the number of selected components was exactly the same as the total number of events, which was suspected to be accidental.

The selected principal component Cox regression (PCR) model was then validated using 10-fold cross validation; a summary of the performance statistics is presented in Table 70. The optimisms are quite large for every single statistics, which was probably due to overfitting.

Table 70. CV Performance for PCR Model – NKI70 Data (Original)

	index.orig	training	Test	optimism	index.corrected
Dxy	-0.9392	-0.9486	-0.6820	-0.2666	-0.6726
R2	0.8138	0.8380	0.4606	0.3774	0.4364
Slope	1.0000	1.0000	0.2208	0.7792	0.2208
D	0.4925	0.5361	0.2738	0.2623	0.2302
U	-0.0046	-0.0052	1.7426	-1.7478	1.7432
Q	0.4971	0.5414	-1.4688	2.0102	-1.5130
g	10.7989	13.2515	2.9586	10.2929	0.5059

Table 71. CV Errors and CV AUCs for PCR Model – NKI70 Data (Original)

Months	Prediction Errors (95% PCI)	Prediction AUCs (95% PCI)
1	0.003 (0.000, 0.006)	0.964 (0.909, 0.985)
2	0.015 (0.006, 0.026)	0.944 (0.884, 0.983)
3	0.013 (0.005, 0.023)	0.937 (0.889, 0.983)
4	0.013 (0.005, 0.024)	0.938 (0.896, 0.981)
5	0.019 (0.008, 0.033)	0.944 (0.902, 0.980)
6	0.016 (0.007, 0.027)	0.957 (0.924, 0.983)
7	0.033 (0.017, 0.054)	0.947 (0.500, 0.985)
8	0.033 (0.017, 0.054)	0.957 (0.931, 0.984)
9	0.152 (0.102, 0.208)	0.964 (0.938, 0.985)
10	0.174 (0.121, 0.232)	0.968 (0.946, 0.985)
11	0.174 (0.121, 0.232)	0.968 (0.946, 0.985)
12	0.212 (0.150, 0.272)	0.957 (0.896, 0.984)
13	0.240 (0.174, 0.303)	0.953 (0.900, 0.984)
14	0.240 (0.174, 0.303)	0.956 (0.904, 0.983)

As previously mentioned, the PCR model was cross validated with the original

NKI70 data; the prediction errors and time-dependent AUCs had to be assessed with the same data set, thus they should also be referred to as the CV errors and CV AUCs. The CV errors are presented in Table 71 and the corresponding plots are displayed in Figure 114. The prediction performance of the model was unexpectedly good in terms of both the CV errors and the CV AUCs, since the performance was evaluated using the same dataset. It was also the main reason of overfitting. Besides the predictor errors, the CV AUCs for the selected PCR models are also summarized in Table 71 and the corresponding plot is displayed in Figure 115.

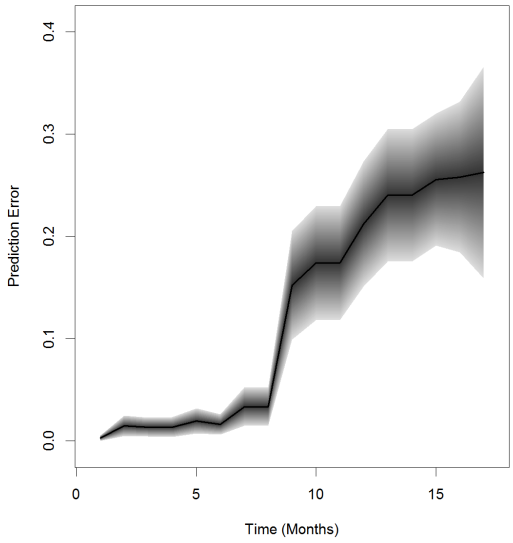


Figure 114. Cross Validation Errors for PCR Model – NKI70 Data (Original)

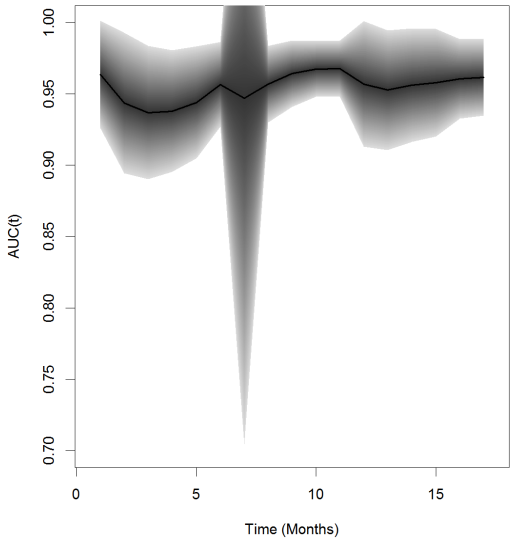


Figure 115. Cross Validation AUC(t) for PCR Model – NKI70 Data (Original)

For prognostic factor detection, it was almost impossible to identify important factors based on the selected PCR model, since it was impractical to read through the 48 by 76 loading matrix.

4.2.1.3.4 Partial Least Squares Cox Regression

For partial least squares Cox regression models, the actual analyses were performed on the PLS components, which was constructed from the original covariates, to achieve the maximum correlation with the survival outcome. For this model, the multicollinearity should be of no concerns, since the PLS components are orthogonal.

However, for this case study, interactions and polynomial transformations were also considered for PLS Cox models, just to achieve a better understanding of this approach.

Similar to the penalized Cox regression models, 3 PLS Cox regression models were performed incorporating the original 76 factors, all 2925 covariate terms (including the original factors and all pair-wise interactions), and incorporating all 735 potential polynomial terms, separately; the models are referred to as PLS Cox linear, PLS Cox interaction and PLS Cox polynomial models, respectively. Once again, all of the models were cross validated with the training set.

4.2.1.3.4.1 Partial Least Squares (PLS) Cox Linear Model

Of 76 factors from the training set, which included 32 metastasis events observed from 108 subjects, initially 25 PLS components were intended for the PLS Cox linear model. The model was cross validated via AIC rule; results are presented in Table 72.

Table 72. CV AIC of PLS Cox Linear Model – NKI70 Data

Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC
0	266.9	5	152.4	10	88.53	14	44.6	18	44.3	22	48.4
1	248.8	6	139.8	11	77.5	15	42.5	19	46.0	23	49.6
2	196.9	7	127.5	12	59.7	16	42.2	20	47.4	24	50.8
3	180.2	8	112.7	13	50.6	17	43.5	21	48.8	25	311.3
4	166.2	9	97.92								

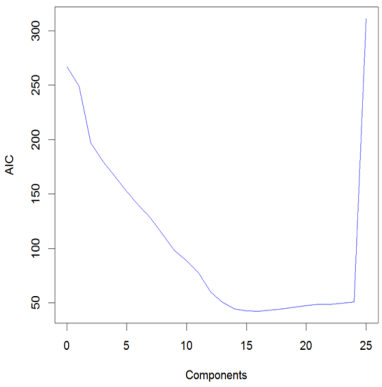


Figure 116. CV Performance of PLS Cox Linear Model – NKI70 Data

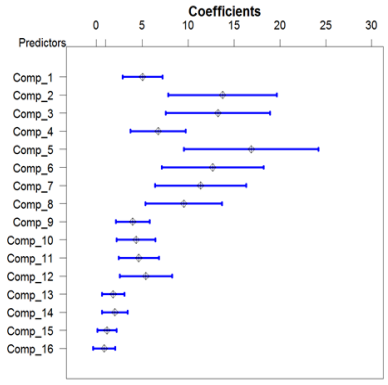


Figure 117. Forest Plot of Component Coefficients from PLS Cox Linear Model – NKI70 Data

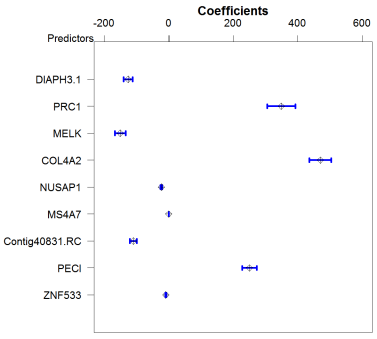


Figure 118. Forest Plot of Coefficients for the Top 9 Factors with the Maximum Loadings from PLS Cox Linear Model – NKI70 Data

The plot of the model AICs against the number of components in the PLS Cox Linear model is displayed in Figure 116. The model achieved the minimum AIC with the first 16

components, then AIC gradually inclined and at the component 25, the model AIC jumped up abruptly. Thus the best PLS Cox linear model should retain the first 16 PLS components.

The coefficients for the first 16 components were estimated from a typical Cox PH model with the 16 PLS components as the covariates; the coefficients corresponding to the 16 PLS components are presented in Table 73; noted that component 16 had p-value of 0.1674, not significant at 0.05 level, but it was still kept in the final model just to be conservative.

Table 73. Component Coefficients from PLS Cox Linear Model – NKI70 Data

Comp	Coef	HR	SE (Coef)	z	P-val
1	5.0367	153.9608	1.1050	4.56	<.0001
2	13.7563	9.4251E+05	3.0229	4.55	<.0001
3	13.2505	5.6835E+05	2.9040	4.56	<.0001
4	6.7358	842.0456	1.5284	4.41	<.0001
5	16.8814	2.1453E+07	3.7367	4.52	<.0001
6	12.6750	3.1967E+05	2.8329	4.47	<.0001
7	11.3521	8.5145E+04	2.5326	4.48	<.0001
8	9.5231	1.3672E+04	2.1223	4.49	<.0001
9	3.9733	53.1588	0.9385	4.23	<.0001
10	4.3393	76.6547	1.0794	4.02	0.0001
11	4.6306	102.5805	1.1152	4.15	0.0000
12	5.3949	220.2728	1.4575	3.70	0.0002
13	1.8371	6.2785	0.6212	2.96	0.0031
14	2.0187	7.5287	0.7096	2.84	0.0044
15	1.1576	3.1822	0.5270	2.20	0.0281
16	0.8565	2.3549	0.6203	1.38	0.1674

For PLS Cox linear model, forest plot for coefficients (the log relative hazard) of the PLS components and the corresponding 95% CIs are presented in Figure 117.

The coefficients for the original 76 factor were obtained by transforming the coefficients of the PLS components using the loading matrix from the PLS Cox linear model, where the loadings reflected the importance of the factors with respect to the survival outcome. Of the 76 factors, the top 9 factors with the most absolute loading from the PLS Cox linear model were selected; the corresponding coefficients are presented in Table 74; the coefficients of all 76 factors are presented in Appendix 17. For this model, each factor should have 16 loadings corresponding to the 16 PLS

components, only the maximum (in absolute values) for each factor is presented. Although there were 16 PLS components, there was only one coefficient corresponding to each of the 76 factors. The forest plot of the coefficients for the top 9 factors with the maximum absolute loadings from the PLS Cox linear model is displayed in Figure 118.

Table 74. Coefficients of the Top 9 Factors with the Maximum Loadings from PLS Cox Linear Model – NKI70 Data

Factors	Coef	SE (Coef)	Loadings
DIAPH3.1	-126.19	-7.155	0.5487
PRC1.1	349.00	22.270	0.4877
MELK	-150.04	-8.310	0.4405
COL4A2	470.21	17.459	0.4137
NUSAP1	-22.44	-1.457	-0.4026
MS4A7	0.11	0.007	-0.3909
Contig40831.RC	-109.37	-5.515	-0.3841
PECI	250.56	11.370	0.3760
ZNF533	-8.77	-0.990	-0.3746

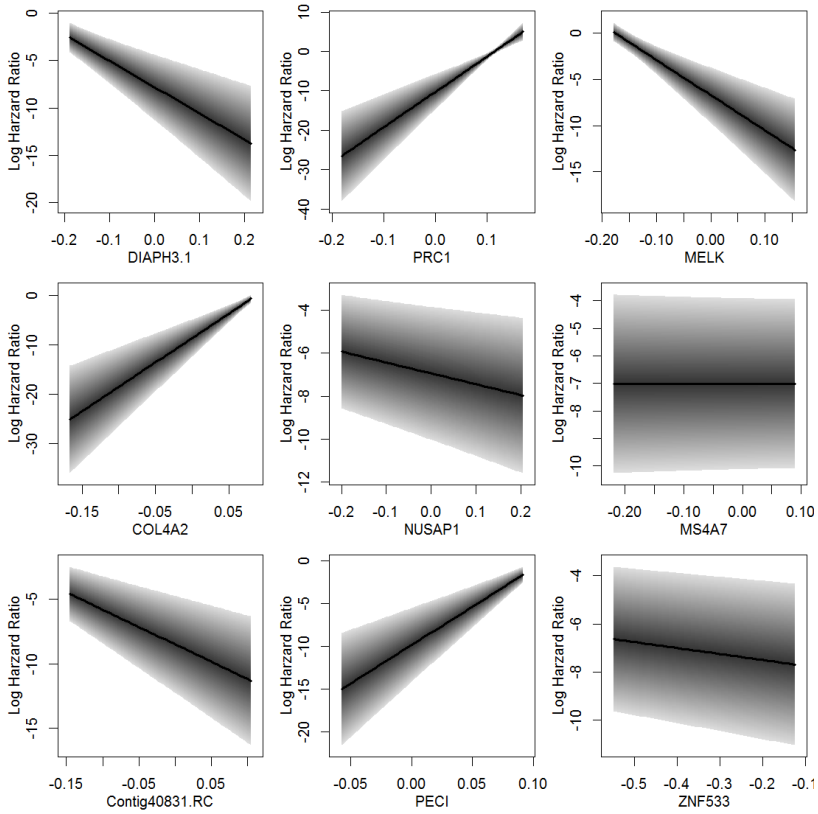


Figure 119. Log HR vs. Factors for the Top 9 Factors with the Most Loadings from the PLS Cox Linear Model – NKI70 Data

The plots of the log relative hazard vs. each of the 9 factors with the maximum absolute loadings from the PLS Cox linear model are displayed in Figure 119. Theoretically, only factors with loadings of 0 are irrelevant; however, in practice, loadings can never reach 0. Thus, a soft threshold was employed to select important factors; all factors above the predetermined soft threshold were considered as important. For example, a total of 11 factors with absolute loadings ≤ 0.15 were considered irrelevant; 45 factors with absolute loadings > 0.15 and ≤ 0.30 were considered relevant; only 20 factors with loadings > 0.3 were considered as important prognostic factors, the 20 factors are ER.Pos, Contig63649.RC, NUSAP1, DIAPH3.1, DIAPH3.2, KNTC2, ZNF533, PECI, Contig40831.RC, TGFB3, MELK, COL4A2, DCK, FBXO31, LOC643008, MS4A7, AP2B1, PITRM1, PRC1 and NM.004702.

The prediction performance of the PLS Cox linear model was then evaluated based on the test set. For cross comparisons, The prediction errors and the time-dependent AUCs of the PLS Cox linear model together with the other PLS Cox models will be presented in Table 81 and Table 82; the corresponding plots will be displayed in Figure 128 and Figure 129, respectively.

4.2.1.3.4.2 PLS Cox Interaction Model

The PLS Cox interaction model was cross validated using the training set with 2925 covariate terms from 108 subjects and 32 events; a total of 30 components were initially intended, cross validation only retrieved the first 25 PLS Components. The model AICs from the cross validation are presented in Table 75.

Table 75. CV Performance of PLS Cox Interaction Model – NKI70 Data

Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC
0	266.9	4	80.2	8	28.1	12	31.9	16	38.9	20	45.4	23	49.5
1	188.5	5	52.4	9	26.2	13	33.0	17	39.6	21	46.2	24	51.2
2	148.3	6	35.1	10	29.6	14	34.5	18	40.4	22	48.6	25	55.5
3	109.7	7	31.7	11	31.4	15	37.9	19	43.0				

Figure 120 displays the AIC against the number of components from the PLS Cox Interaction model. The model reached the minimum AIC with the first 9 components; the components beyond component 9 should be excluded to construct the best PLS Cox interaction model.

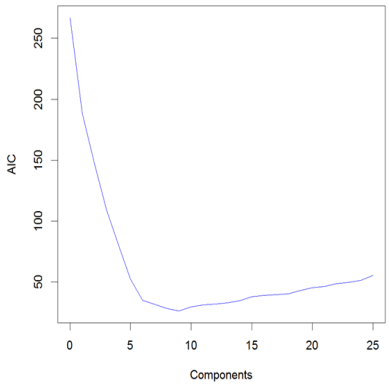


Figure 120. CV Performance of PLS Cox Interaction Model – NKI70 Data

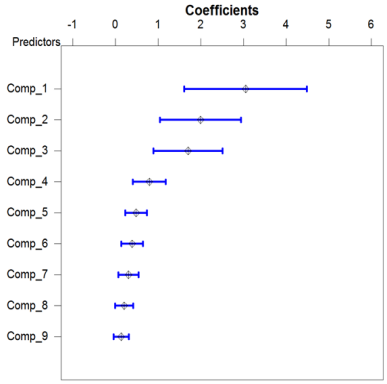


Figure 121. Forest Plot of the 9 PLS Component from PLS Cox Interaction Model – NKI70 Data

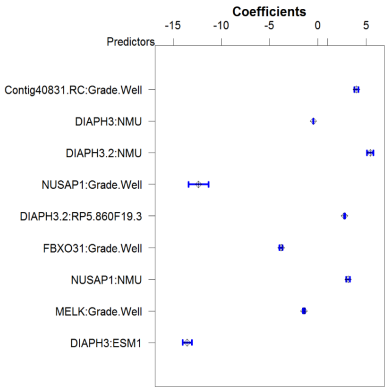


Figure 122. Forest Plot of the Top 9 Factors with the Maximum Loadings from the PLS Cox Interaction Model – NKI70 Data

The coefficients of the 9 PLS components were estimated from the PLS Cox interaction model with the training set; results are presented in Table 76; component 9 was not significant (P-value = 0.1530) and component 8 was boundary significant (P-value= 0.0599), both of them were still kept in the final model, just to be conservative. Forest plot of the coefficients for the 9 PLS components is presented in Figure 121; the coefficients were used instead of the hazard ratios, since the 95% CIs for hazard ratios were too diverse.

Table 76. Component Coefficients from PLS Cox Interaction Model – NKI70 Data

Comp	Coef	HR	SE (Coef)	z	Pr(>|z|)
1	3.0475	21.0634	0.7318	4.16	0.0000
2	1.9931	7.3382	0.4853	4.11	0.0000
3	1.6998	5.4731	0.4126	4.12	0.0000
4	0.7907	2.2049	0.1980	3.99	0.0001
5	0.4831	1.6211	0.1303	3.71	0.0002
6	0.3894	1.4761	0.1290	3.02	0.0025
7	0.3030	1.3539	0.1222	2.48	0.0131
8	0.2026	1.2246	0.1077	1.88	0.0599
9	0.1319	1.1410	0.0923	1.43	0.1530

With the loading matrix from the PLS Cox interaction model, the coefficients of the original 2925 covariates were obtained by inverse transformation from the coefficients of the PLS components; full summary of all 2925 coefficients is not presented; only the coefficients of the top 9 factors with the max absolute are presented in Table 77. The

forest plot of the coefficients for the top 9 factors with the max absolute loadings from the PLS Cox interaction model is displayed in Figure 122.

Table 77. Coefficients of the Top 9 Factors with the Maximum Loadings from PLS Cox Interaction Model – NKI70 Data

	Coef	SE (Coef)	Loadings		Coef	SE (Coef)	Loadings
Contig40831.RC:Grade.Well	4.0	0.11	0.31	FBXO31:Grade.Well	-3.8	-0.08	0.14
DIAPH3:NMU	-0.5	-0.02	0.29	NUSAP1:NMU	3.1	0.1	0.14
DIAPH3.2:NMU	5.4	0.16	0.19	MELK:Grade.Well	-1.4	-0.05	0.13
NUSAP1:Grade.Well	-12.4	-0.52	0.17	DIAPH3:ESM1	-13.5	-0.24	0.12
DIAPH3.2:RP5.860F19.3	2.7	0.03	0.15				

The plots of log relative hazard vs. each of the 9 factors with the max absolute loadings from the PLS Cox interaction model are displayed in Figure 123.

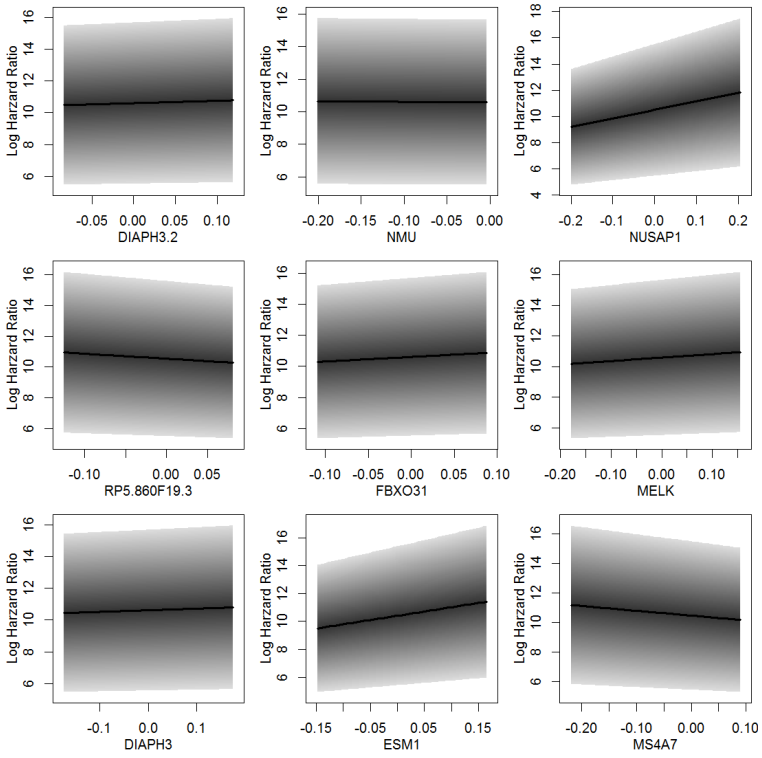


Figure 123. Log HR vs. Factors for the Top 9 Factors with the Most Loadings from the PLS Cox Interaction Model – NKI70 Data

The prediction performance of the PLS Cox interaction model was evaluated based on the test set. For cross comparisons with other PLS Cox models, the prediction errors and the time-dependent AUCs will be presented in Table 81, Table 82, respectively and

the corresponding plots will be displayed in Figure 128 and Figure 129, respectively.

4.2.1.3.4.3 PLS Cox Polynomial Model

The PLS Cox polynomial model with all 735 covariates was initially intended with 25 components. However, the majority of these covariates were probably not relevant to the survival outcomes. Thus, the coefficients for most of the covariates were close to 0 or infinity; as a result, for the PLS components, many entries of the transformation matrix were close to infinity or zero; which had made the PLS Cox polynomial model difficult to convergent. To improve the convergence, all original factors were normalized first; additionally, instead of keeping all none-zero coefficients, only significant components (p-value <0.20) were kept in the final model. Moreover, to further improve the efficiency, all covariate terms were normalized before fitting to the PLS Cox polynomial model. With the above modification, the PLS Cox polynomial model was able to converge and the AICs for the PLS Cox polynomial model are presented in Table 78.

Table 78. CV Performance of PLS Cox Polynomial Model – NKI70 Data

Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC	Comp	AIC
0	266.9	2	162.4	4	120.0	6	71.4	8	39.00	10	25.9
1	195.3	3	139.7	5	90.3	7	48.4	9	33.2		

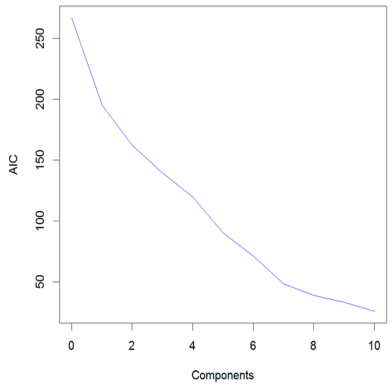


Figure 124. CV Performance of PLS Cox Polynomial Model – NKI70 Data

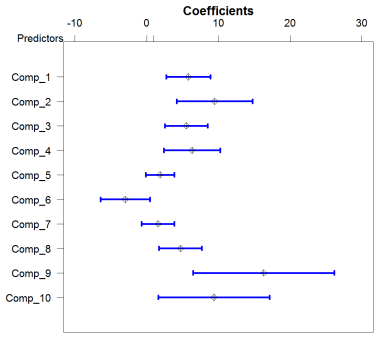


Figure 125. Forest Plot of the 10 PLS Components for PLS Cox Polynomial Model – NKI70 Data

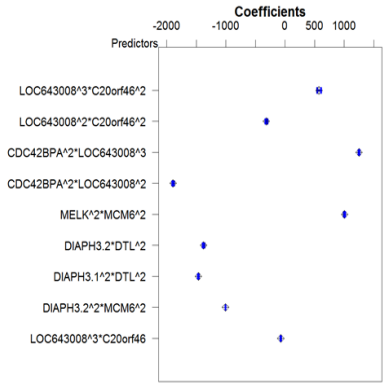


Figure 126. Forest Plot of the 9 Covariates with the Maximum Loadings from PLS Cox Polynomial Model – NKI70 Data

Figure 124 displays the AIC against the number of components in the model. As noted, the AIC curve showed slightly different patterns from the one for PLS Cox linear model (Figure 116) or PLS Cox interaction model (Figure 120). The CV of the PLS Cox

polynomial model was only able to achieve10 PLS components and the model reached the minimum AIC with the first 10 components; therefore best PLS Cox polynomial model should include the first 10 PLS components.

The coefficients of the first 10 components are presented in Table 79; note that component 7 was not significant at the 0.05 level (p-value = 0.1735), but it was still kept in the final model, since it was needed to derive components 8, 9 and 10, subsequently. (For PLS Cox models, the order of the components are important, since higher ordered components are derived from lower ordered components.) The forest plot of the coefficients of the 10 PLS Components is displayed in Figure 125.

Table 79. Component Coefficients from PLS Cox Polynomial Model – NKI70 Data

Comp	Coef	HR	SE (Coef)	z	P-val
1	5.7971	3.2934E+02	1.5727	3.69	0.0002
2	9.4679	1.2938E+04	2.6986	3.51	0.0005
3	5.5114	2.4749E+02	1.5166	3.63	0.0003
4	6.3110	5.5061E+02	2.0102	3.14	0.0017
5	1.8380	6.2837E+00	1.0146	1.81	0.0701
6	-2.9926	5.0155E-02	1.7538	-1.71	0.0879
7	1.5778	4.8443E+00	1.1593	1.36	0.1735
8	4.6961	1.0952E+02	1.5174	3.09	0.0020
9	16.3070	1.2079E+07	5.0164	3.25	0.0012
10	9.3612	1.1628E+04	3.9528	2.37	0.0179

The coefficients for the original 735 covariate terms were obtained by back-transforming the coefficients of the PLS components with the loading matrix from the PLS Cox polynomial model. Table 80 presents the coefficients of the top 9 covariate terms with the max absolute loadings from the PLS Cox polynomial model.

Table 80. Coefficients of the Top 9 Covariate Terms with the Maximum Loadings from PLS Cox Polynomial Model – NKI70 Data

Covariate Terms	Coef	SE (Coef)	Loading	Covariate Terms	Coef	SE (Coef)	Loading
LOC643008^3:C20orf46^2	574.9	16.99	0.83	DIAPH3.2:DTL^2	-1377.4	-6.56	0.42
LOC643008^2:C20orf46^2	-316.5	-8.65	0.67	DIAPH3.1^2:DTL^2	-1463.4	-4.67	0.37
CDC42BPA^2:LOC643008^3	1250.8	4.66	0.48	DIAPH3.2^2:MCM6^2	-1007.0	-1.00	0.36
CDC42BPA^2:LOC643008^2	-1892.1	-7.39	0.44	LOC643008^3:C20orf46	-72.4	-2.26	0.32
MELK^2:MCM6^2	1006.0	5.68	0.43				

The coefficient estimates for all 735 covariates are presented in Appendix 18. Again each covariate should have 10 loadings corresponding to the 10 PLS components; only

the max absolute loading is selected out of the 10 loadings for each covariate and presented in Table 80. Although there were 10 loadings for each covariate term, after back-transformation from the coefficients of the PLS components using the loading matrix, only one coefficient was estimated corresponding to each covariate terms. The forest plot of the coefficients for the top 9 covariates with the max absolute loadings is displayed in Figure 118.

The 9 covariate terms were made up of 9 factors, LOC643008, C20orf46, CDC42BPA, LOC643008, MELK, MCM6, DIAPH3.2, DTL and DIAPH3.1; for each of these 9 factors, predictions were made based on the cross validated PLS Cox polynomial model while fixing the rest of the 75 factors (continuous factors were fixed at the medians and the categorical factors were fixed at the most frequent category level); Figure 127 displays the log hazard ratio vs each of the 9 factors.

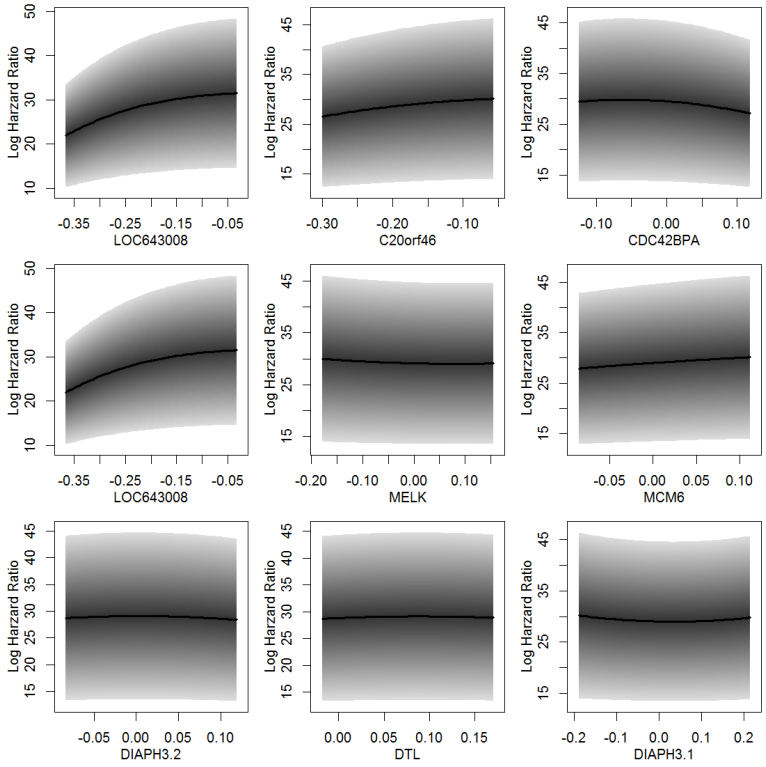


Figure 127. Log HR vs. Factors for the Top 9 Factors with the Most Loadings from the PLS Cox Polynomial Model – NKI70 Data

For this model, the 10 PLS components were constructed from all of the 735 covariate terms, it was impossible to manually check the loading matrix; again, a soft threshold was employed to select the important covariates. A total of 685 covariates with

absolute loading ≤ 0.15 were considered irrelevant; 39 covariates with loading > 0.15 and ≤ 0.30 were considered as mild relevant; only 11 terms with loading > 0.3 were considered as important, the 11 terms are DIAPH3.2^2:MCM6^2, DIAPH3.1^2:DTL^2, LOC643008^2:C20orf46^2, LOC643008^3:C20orf46, DIAPH3.2:DTL^2, CDC42BPA^2:LOC6430082, MELK^2:PITRM1^3, LOC643008^3:C20orf46^2, QSCN6L1:DIAPH3.2^2, , MELK^2:MCM6^2, CDC42BPA^2:LOC643008^3.

4.2.1.3.4.4 Prediction Performance Comparison of Intended Survival Models for the Real Word Case Study (NKI70 Data)

For PLS Cox models, the prediction errors and corresponding 95% PCIs are presented Table 81 and the prediction errors of the 3 PLS Cox models are superimposed on top of each other in Figure 128. Of the 3 models, PLS polynomial model had the best prediction errors except that it became slightly worse than the PLS interaction model in the tail; the PLS interaction model had similar prediction errors with the PLS linear model for the first 9 months, but it became the best of the 3 models after that; the PLS linear model had almost the worst prediction errors. However, the differences were very minimal; the maximum difference was less than 0.045 numerically.

Table 81. Prediction Errors of PLS Cox Models – NKI70 Test Set

Months	PLS Linear (95% PCI)	PLS Interaction (95% PCI)	PLS Polynomial (95% PCI)
1	0.051 (0.000, 0.139)	0.058 (0.000, 0.153)	0.051 (0.000, 0.139)
2	0.109 (0.028, 0.222)	0.117 (0.028, 0.225)	0.109 (0.028, 0.222)
3	0.214 (0.083, 0.361)	0.210 (0.083, 0.361)	0.196 (0.083, 0.333)
4	0.265 (0.135, 0.417)	0.261 (0.091, 0.443)	0.247 (0.111, 0.411)
5	0.345 (0.167, 0.528)	0.346 (0.150, 0.549)	0.327 (0.167, 0.505)
6	0.345 (0.167, 0.528)	0.346 (0.150, 0.549)	0.327 (0.167, 0.505)
7	0.356 (0.181, 0.556)	0.367 (0.151, 0.569)	0.334 (0.148, 0.532)
8	0.356 (0.181, 0.556)	0.367 (0.151, 0.569)	0.334 (0.148, 0.532)
9	0.405 (0.214, 0.611)	0.344 (0.132, 0.542)	0.360 (0.189, 0.517)
10	0.405 (0.214, 0.611)	0.344 (0.132, 0.542)	0.360 (0.189, 0.517)
11	0.405 (0.214, 0.611)	0.344 (0.132, 0.542)	0.360 (0.189, 0.517)
12	0.403 (0.214, 0.611)	0.342 (0.131, 0.545)	0.359 (0.187, 0.520)
13	0.397 (0.214, 0.583)	0.338 (0.137, 0.531)	0.357 (0.170, 0.527)
14	0.397 (0.214, 0.583)	0.338 (0.137, 0.531)	0.357 (0.170, 0.527)

The time-dependent AUCs and the corresponding 95% PCIs are presented in Table 82 and the plots of the time-dependent AUCs for the 3 models are displayed in Figure

131. In terms of the time-dependent AUCS, the PLS Cox interaction model had the best performance and the PLS Cox linear model had the worst performance; the AUC curve of the three model were almost parallel with each other.

Table 82. Prediction AUCs of PLS Cox Models – NKI70 Test Set

Months	PLS Linear (95% PCI)	PLS Interaction (95% PCI)	PLS Polynomial (95% PCI)
1	0.562 (0.500, 1.000)	0.694 (0.500, 1.000)	0.645 (0.500, 1.000)
2	0.534 (0.500, 0.777)	0.681 (0.500, 0.857)	0.628 (0.500, 0.832)
3	0.540 (0.500, 0.777)	0.712 (0.500, 0.860)	0.635 (0.500, 0.834)
4	0.537 (0.500, 0.774)	0.701 (0.500, 0.856)	0.622 (0.500, 0.833)
5	0.529 (0.500, 0.776)	0.676 (0.500, 0.857)	0.610 (0.500, 0.832)
6	0.516 (0.500, 0.760)	0.643 (0.500, 0.855)	0.597 (0.500, 0.831)
7	0.513 (0.500, 0.745)	0.645 (0.500, 0.859)	0.618 (0.500, 0.842)
8	0.514 (0.500, 0.739)	0.647 (0.500, 0.859)	0.623 (0.500, 0.841)
9	0.516 (0.500, 0.744)	0.653 (0.500, 0.859)	0.631 (0.500, 0.841)
10	0.519 (0.500, 0.761)	0.653 (0.500, 0.860)	0.634 (0.500, 0.839)
11	0.517 (0.500, 0.750)	0.652 (0.500, 0.858)	0.638 (0.500, 0.840)
12	0.517 (0.500, 0.756)	0.679 (0.500, 0.859)	0.650 (0.500, 0.847)
13	0.523 (0.500, 0.769)	0.691 (0.500, 0.860)	0.650 (0.500, 0.844)
14	0.523 (0.500, 0.769)	0.691 (0.500, 0.860)	0.650 (0.500, 0.844)

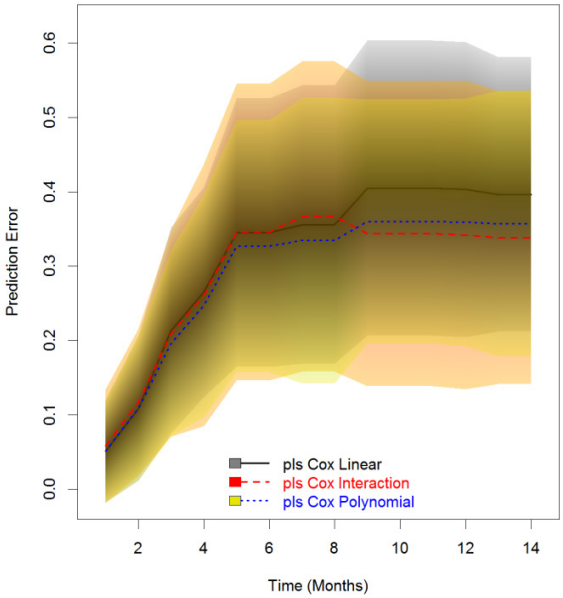


Figure 128. Prediction Errors of PLS Cox Models – NKI70 Test Set

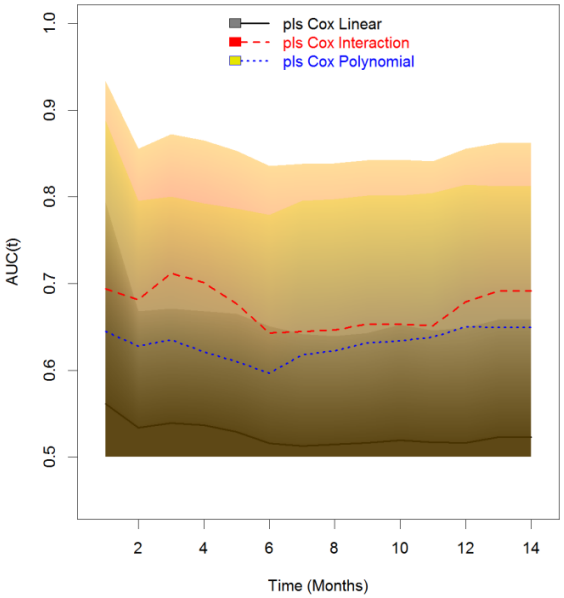


Figure 129. Prediction AUCs of PLS Cox Models – NKI70 Test Set

Comparing the log HR plots from the 3 PLS Cox models, the PLS Cox linear model had the smallest interval of the three PLS Cox models, which suggests that the PLS Cox

linear model was the most powerful for making inferences about the coefficients of the factors; however the prediction performance was almost the worst of the three models, which was probably due to underfitting.

4.2.1.4 Result Summary of the Case Study

For cross comparison of all survival models intended for this case study, the prediction errors for all models are displayed in Figure 130 and the time-dependent AUCs for all models are displayed in Figure 131; the 95% PCIs are not presented. As noted in section 4.2.1.3.3, principal component Cox regression (PCR) were trained with all 144 subjects of the original NKI70 data, therefore it had the almost perfect performance in terms of both prediction errors and time-dependent AUCs, except that it was only slightly worse than ridge linear, elastic-net linear and elastic-net polynomial Cox models at the tail (after month 11). However, it was unfair to compare this model with the rest of the models, since the same dataset with all available subjects (144 subjects) was used for both training and testing (assessing the prediction performance) for the PCR model, while all the other models were trained and cross validated with the training set (108 subjects) and prediction performance measurements, including prediction errors and time-dependent AUCs, were assessed with the test set.

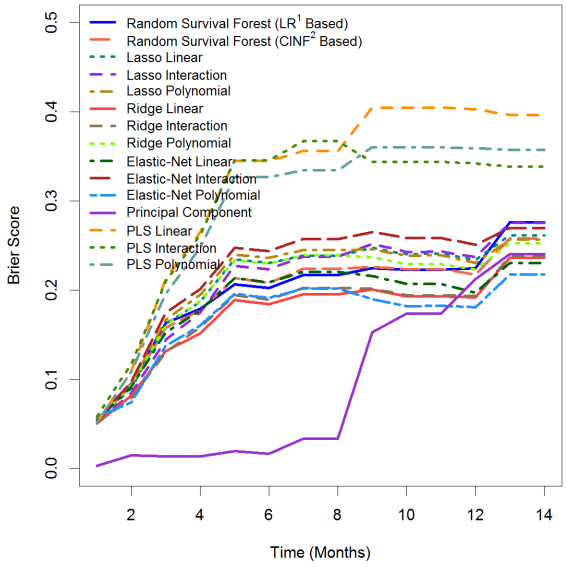


Figure 130. Prediction Errors of All Models – NKI70 Data

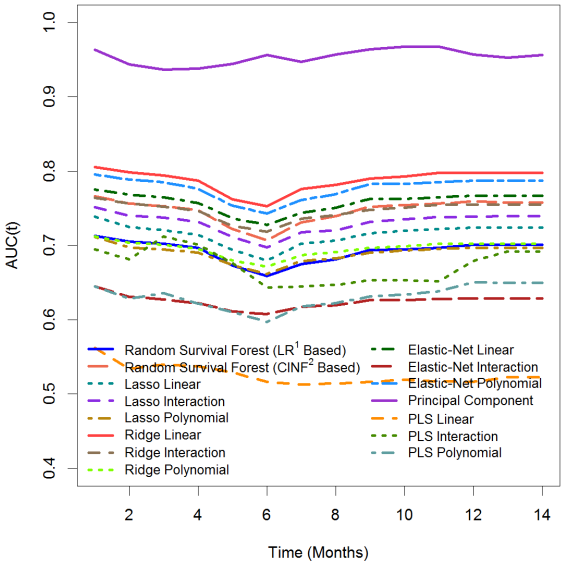


Figure 131. Time-Dependent AUC of All Models – NKI70 Data

Of the rest of the survival models (excluding the PCR model), ridge Cox linear, ridge

Cox interaction, elastic-net Cox linear and elastic-net Cox polynomial models had the best prediction errors. The 3 PLS Cox models (linear, interaction and polynomial) had the worst prediction errors.

While the prediction AUCs were almost parallel among all different models. The PCR model had the best AUCs again; otherwise, the ridge Cox linear, elastic-net Cox polynomial, elastic-net Cox linear had extraordinary prediction AUCs, followed by ridge interaction, conditional inference based RSF and lasso interaction, lasso linear and ridge Cox polynomial models; the log-rank based RSF model had moderate AUCs comparing with the other semi-parametric models. The elastic-net Cox interaction as well as the three PLS Cox models had the worst prediction AUCs; as mentioned previously, the prediction AUCs of the elastic-net Cox interaction model and the PLS Cox linear model was extremely disappointing due to significant underfitting.

While comparing the PLS models with all other survival models, the prediction performance was unexpectedly unsatisfying, which was possibly due to the system errors as discussed in section 4.1.2.3.7. Additionally, unlike the PLS Cox interaction and PLS Cox linear models, the PLS Cox polynomial model could not achieve convergence with all covariates in their original scale, while the same model was able to achieve convergence after all covariate terms were normalized; however, normalization of all covariate terms could have inflated the noise (covariates) for the model. This was probably the reason why the prediction performance of the PLS Cox polynomial model was worse than the PLS Cox interaction model.

In terms of model selection, the elastic-net interaction model selected the minimum number of covariates, however the lasso-like penalization parameters had caused significant underfit; otherwise, the 3 lasso Cox models selected the smallest number of covariates; the elastic-net Cox polynomial and linear models were also efficient for selecting important prognostic factors. In terms of prediction future events, ridge linear, elastic-net polynomial and elastic-net linear Cox models had the best predictions performance; the ridge Cox interaction model was only slightly worse. If both purposes (prediction of future event and prognostic factors detection) were of equally importance, elastic-net Cox linear and elastic Cox polynomial models were the most effective tools. For the two nonparametric RSF models, the prediction errors of the log-rank based RSF

and conditional inference based RSF models were moderate; however, in terms of prediction AUCS, the conditional inference based RSF model was better than expected, comparable to the ridge interaction model, while the AUCs of the log-rank based RSF was still moderate; additionally, the two RSF models were very effective for prognostic factor detections.

For the two derived input Cox regression models, the PCR and PLS Cox models have limited capabilities for selecting prognostic factors due to the construction of latent components. The prediction performance of the PCR model for this case study was not comparable to the other models since the same dataset was use for both training and testing, thus the performance of the model could not be generalized to other studies; additionally, the model may still be non-estimable for too many components. Unlike the PCR model, the PLS models do not construct as many components, thus it is unlikely to encounter a non-estimable PLS model; however the PLS Cox model had unsatisfying prediction performance, which was possibly due to the system errors involved in calculating the predictive survival probability. Thus, the two derived input Cox regression models are not recommended unless no other choices are available.

Chapter 5. Conclusions and Discussions

In this research paper, 3 typical Cox PH models were evaluated, including Cox linear model, Cox model with *RCS* transformations, Cox model with *FP* transformations; two nonparametric RSF models were studied, including log-rank based random survival forest and conditional inference based survival forest; multiple penalized Cox regression models were assessed extensively, including lasso, ridge and elastic-net Cox regression models; and two derived input Cox regression model including principal component Cox regression (PCR) and partial least squares (PLS) Cox regression models were also implemented.

Concerning prognostic factors detection, the Cox linear model and Cox models with nonlinear transformations are semi-parametric approaches, they are the most stringent of all intended survival models, however they can provide unbiased parameter estimates and prediction performance for these models are better than most survival models; however they do not work when the number of factor is close to or more than the total number of events. Of the 3 typical Cox PH models, the Cox model with *FP* transformation had the best prediction performance, possibly due to the inclusion of time-dependent treatment effect; the Cox model with *RCS* transformation was not very sensitive to non-proportionality, the prediction performance was not as good as the one with *FP* transformation; the Cox PH linear model was the simplest of the three typical Cox models, but the prediction performance of the model was quite satisfactory. Additionally, typical Cox PH models are well developed with many options and extensions; the recurrent event extensions on top of the Cox PH model are very useful, the extension Cox PH models can deal with multiple event survival data, such as event recurrence or competing events survival outcomes. In addition, typical Cox PH models are also the most popular survival models; almost all features are well studied, therefore they are the most convenient solutions for most survival problems.

The two random survival forest models including log-rank based RSF (LR-RSF) and conditional inference based RSF models (CINF-RSF) are the most flexible of all intended survival models; they are nonparametric, thus they have inherited all the flexibility of a nonparametric approach. They are developed with no or very little model assumptions; there are no concerns of non-proportionalities, multicollinearity or nonlinearity. In the

simulation study, both RSF models detected all of the prognostic factors as those detected by the typical Cox regression models and they had moderate prediction performance, slightly worse than typical Cox PH models. Of the two random survival forest approaches, the prediction performance of the LR-RSF model were similar to the CINF-RSF model in the simulation study; but in the real world case study, the prediction performance of the LR-RSF was consistently worse than the CINF-RSF model; the reason was possibly due to the high correlations among factors. In the real word case study, CINF-RSF model had better prediction performance, since this model builds forest trees based on conditional probabilities; thus the CINF-RSF model works better for highly correlated survival data.

In this research paper, three penalized Cox regression models including lasso Cox regression, ridge Cox regression and elastic-net Cox regression models were studied and 3 options including linear, interaction and polynomial transformations were assessed intensively. During the study, it was found that the cross validation via partial log likelihood option was not very stable. In which, the selection of the penalization term, λ, was based on partial log likelihood deviance, which was not a good reflection of the model performance. In fact, the deviance is related to the joint likelihood of the full model, as derived by Simon et al. (2011)[154], the maximization of the joint likelihood has become the minimization process for the partial log likelihood deviance measurement. But both the deviance and the full model likelihood are dependent upon the initial Cox model with all covariate terms. It basically assumes that the full model has the best fit of the survival data, which may not be necessarily true. Instead, during the studies, it was found that the deviance was only related to the active survival data through the full model fit, and the search of the penalization parameter was neither stable nor robust, it was very much depending on the seed chosen; additionally, the deviance measurement might have multiple local minimum, all searching algorithms should be able to locate a local minimum, but there was no guaranteed to find the global minimum. Therefore, searching for the penalization parameters could be problematic.

Thus, a modified cross validation process was developed; the cross validation was carried out using Brier Scores as the selection rules, which was proved to be very stable and robust; however for lasso and ridge Cox regression models, this modified cross

validation algorithm could not take more factors than the actual number of observations in the dataset. While for elastic-net Cox regression models, an interval search algorithm was deployed within the CV process, to achieve reasonably robust estimates of the penalization terms (α and λ) simultaneously. This modified CV algorithm did not guarantee to find the global minimum of the partial log likelihood deviance, but it guaranteed to find the model with the minimum cross validation errors given the training set.

Besides, two derived input Cox regression models, principal component and PLS Cox models, were also studied. The prediction performance of the PCR model was similar to the typical Cox models (except for the Cox model with *FP* transformation), mainly due to the components were constructed by most or all of the covariate terms. The approach is consisted of two steps. The first step is to construct the principal components independently from the survival outcomes, thus it has been widely used for variable reduction. The second step is to perform typical Cox regression analysis over the constructed components. Hence, this approach should have the same advantage as the typical Cox PH models, such as unbiased estimator (for components), and capability of dealing with recurrent event or competing event survival data. Additionally, since the constructed components are orthogonal, the model can handle more components than typical Cox models can deal with covariates. However, aside from the advantages, the model is limited for prognostic factor detections, and the results are difficult to interpret since the coefficients are estimated corresponding to the latent components which has to be further transformed back for the original factors for interpretation; additionally, when the constructed components are more than the total number of events, the model may still be non-estimable due to singularity.

Similarly, the PLS model also derives latent orthogonal components, which are constructed to have the maximum correlation with the survival outcome; therefore multicollinearity should be of no concerned for this analysis. In contrast to the principal component regression which constructs the components to achieve high covariance or correlations within the independent variables (Stone and Brooks, 1990; Frank and Friedman, 1993), the partial least squares regression model derives the components in the directions that can achieve the maximum variance and correlations with the response; the

path of deriving the PLS components is a nonlinear transformation of both the independent and dependent variables. Additionally, unlike the principal component regression, the partial least squares regression model rarely needs to construct more than 30 components; the design matrix of the PLS model should usually be invertible, thus the PLS model is often estimable with the available data. The partial least squares approach, was initially proposed for continuous outcomes, and later extends to categorical outcomes. In this paper, the approach has been generalized to survival outcomes. While this approach has not been widely utilized, and there have been very little publications on this topic. In this research, the partial least squares Cox regression model was implemented and evaluated over two studies, one simulation study and one real world case study on macro array survival data.

The PLS Cox model is implemented using latent components to link the covariates with the survival outcomes, which may not be intuitive for interpretation; in addition, the same factors or covariates should make contributions to all of the PLS components, therefore it could be difficult to identify the important factors or covariates corresponding to the selected PLS components. On the other hand, the PLS components can be considered as uncollected or unseen latent variables from the original dataset. Unfortunately, the prediction performance of the PLS Cox model is not as good as most of the intended survival models due to the unavoidable system errors.

5.1 Findings from the Simulation Study

Of all intended survival models, the Cox model with *FP* transformation had the best prediction performance (including both prediction errors and time-dependent AUCs) at the beginning of the study, which was due to the inclusion of time-dependent treatment effect; but it caught up with the rest of the models at the tail. The prediction performance of the Cox PH linear model was the second of the 3 typical Cox PH models and it exceeded the Cox model with *FP* transformation slightly in the tails. Surprisingly, the prediction performance of the Cox model with *RCS* transformations was the worst of the 3 models, even though the difference was very minimal. This was possibly due to the fact that the model missed the factor Age but falsely selected BMI as one of the important prognostic factors.

For the 2 nonparametric random survival forest models, the prediction performances of the 2 RSF models were reasonably satisfactory; furthermore, the 2 approaches were very convenient since they had little or no model assumptions, such as multicollinearity, proportionality, nonlinearity or interactions etc. Of the two RSF models, both models had similar prediction errors at the beginning of the study; while during the middle of the study, the conditional inference based RSF model had slightly better prediction errors; and the log-rank based RSF behaved slightly better in the tail. In terms of time-dependent AUCs, the log-rank based RSF model was slightly better at the beginning of the study; otherwise the two RSF models were similar. During the analysis, it was found that the two RSF models had some disadvantages; they were black-box approaches, which made them difficult to interpret the mechanism for model selections; the approaches were based on bootstrap aggregation, therefore they could be very resource-consuming for training and cross validation the models and it was extremely time consuming for evaluating the prediction performance.

Of all 3 penalized Cox models, the elastic-net Cox model via interval search had the best time-dependent AUCs. The ridge Cox model was the next by keeping all prognostic factors or covariate terms; and the lasso Cox model was the last. While in terms of prediction errors, the three penalized Cox models were very similar to each other.

The prediction performance of the principal component Cox regression model was above average for the simulation study. The prediction error of this model was excellent, but the prediction AUC was just about average. For PLS Cox model, the prediction performance was almost the worst of all intended models.

In terms of prediction errors, the Cox model with *FP* transformation was the best at the beginning, then it caught up with the rest of models in the tails; the Cox PH linear model, the principal component Cox model and conditional inference based RSF model was the second best; then the Cox model with *RCS* transformation and elastic-net Cox model was the next followed by the ridge, PLS and lasso Cox models. The prediction error of the log-rank based RSF model was the worst during the middle of the study, between 5 to 10 years, otherwise it was comparable to all other models.

In terms of prediction AUCs, the Cox model with *FP* transformation was the best at all time-points; the elastic-net Cox model was the second, followed by Cox PH linear

model, ridge, Cox model with *RCS* transformation and principal component Cox regression model; the lasso and PLS Cox models were the worst. The conditional inference based and log-rank based RSF models were in the middle; however, consider the convenience of the two models due to the nonparametric nature; they can be used as alternative tools for studying survival data.

5.2 Findings from the Real World Case Study on NKI70 Data

For the real world case study on NKI70 data, the conditional inference based RSF had better prediction AUCs than the log-rank based RSF, which was possibly due to the high correlation among factors; since the conditional inference based RSF model was developed based on conditional probabilities, which was specifically good at dealing with correlated data; while the log-rank based RSF model completely discard the multicollinearity among the factors, which might have some impact on the prediction performance. While in terms of prediction errors, the two RSF models were almost equivalent.

For the case study on NKI70 data, the principal component Cox regression (PCR) approach constructed orthogonal latent components, but the model was not estimable with the training set, instead the original NKI70 data was used for both training and testing the PCR model, thus this model had exceptional prediction performance, which could not be generalized to other studies; in addition, it was found that the number of selected principal components was exactly the same as the total number of available events, which was possibly due to a random chance. Additionally, the results of the analysis are not intuitively interpretable because the model was built on the latent components.

Other than the PCR model, the ridge Cox linear model had very good prediction performance; the elastic-net Cox polynomial model also achieved excellent performances (for both prediction errors and prediction AUCs); the elastic-net Cox linear model should be ranked next, followed by the ridge Cox interaction model.

For lasso Cox models, the prediction errors of the 3 lasso Cox models were similar to each other; for time-dependent AUCs, the lasso Cox interaction model had the best AUCs, followed by the lasso Cox linear model; the lasso Cox polynomial model was the worst of 3, which suggested that polynomial transformations were not necessary for lasso

Cox models. Comparing with the other penalized Cox models, the 3 lasso Cox models were almost the worst of all intended penalized Cox models. In general, the lasso Cox models were efficient for selecting prognostic factors, but they should not be used for predictions.

For ridge Cox models, the ridge Cox linear and ridge Cox interaction models had similar prediction errors and prediction errors of the ridge Cox polynomial model was consistently the worst of the 3 options. In terms of time-dependent AUCs, the ridge Cox linear model had the best AUCs and the ridge Cox polynomial model had the worst prediction AUCs, which suggested that the polynomial transformation was not necessary for the ridge Cox models. Additionally, it was also found that the ridge Cox model did not perform prognostic factor selections, the models kept all covariates with the intension to achieve good performance.

For elastic-net Cox models, the elastic-net Cox interaction model had the worst performance (in both prediction errors and AUCs) due to the significant underfit from the lasso-like penalization parameters. The elastic-net Cox linear and elastic-net Cox polynomial models had similar prediction errors; while the prediction AUCs of the elastic-net Cox polynomial model was slightly better than the elastic-net Cox linear model. In general, the elastic-net Cox models were effective for selecting prognostic factors and prediction of unseen outcomes with excellent prediction performance.

For the partial least squares models, the prediction performance was very unsatisfying, possibly due to the system errors as discussed in section 4.1.2.3.7. Additionally, it was found that the PLS Cox polynomial model was only able to achieve convergence unless all covariate terms were normalized; but normalization of all covariate terms should have inflated the noise (covariates) for the model. This was probably the reason why the prediction performance of the PLS Cox polynomial model was worse than the PLS Cox interaction model.

For this study, polynomial transformation did not improve the model performance for lasso and ridge Cox models and slight improvement was observed from the elastic-net Cox polynomial model and PLS Cox polynomial model; pair-wise interaction terms did not improve the performance for ridge and elastic-net Cox model, but slight improvement was observed for lasso Cox interaction and PLS Cox interaction models.

Therefore for the case study, polynomial transformation was unnecessary for lasso and ridge Cox regression models, instead it cost more damage than improvement to the prediction performance of the models; on the contrary, for the elastic-net Cox regression model and partial least squares Cox model, the polynomial transformation achieved slight improvement over the linear ones in the prediction performance.

For lasso Cox regression model, cross validation randomly selected only one term from the correlated group of covariates (note that a polynomial transformation of a particular factor should include multiple terms involving the same factor, the correlation within multiple terms of the same factor should have strong correlations for sure); however elastic-net Cox model was able to select multiple covariates within a correlated group if one was selected from the same group, therefore it should be able to pick up the correlated covariate terms appropriately. For ridge Cox regression, the under-performance for the model with polynomial transformation was possibly due to the same reason for the lasso Cox polynomial model, except that the ridge Cox regression tend to shrink the extra correlated terms to zero instead of dropping them.

5.3 Additional Comments

In this paper, several typical Cox regression models were implemented over the simulation study; but none of them worked with the NKI70 microarray survival data from the real world case study, since the number of covariates was more than the number of events available. They are typical statistical models; therefore if they are able to fit to the data, they should provide unbiased estimates and the prediction performance of these models should be better than most of the survival models. Additionally, the typical Cox regression model has been well developed with many useful options and features; the recurrent event extension of the typical Cox regression model is capable of handling recurrent-event survival data, competing risk survival data, and interval-censored data. As shown in the simulation study, the time-varying treatment effect was handled as if the treatment switching was competing with death, or as if multiple observations occurred within a subject, subjects who switched treatment were considered as having two different observations, except that only one event occurred. In the simulation study, Andersen-Gill (AG) extension of Cox PH model was employed, where different observations from the same subject were considered independent; this assumption was

quite stringent, and it might not be reasonable. Other models such as WLW marginal Cox regression model or PWP conditional Cox regression model are also available, which have more relaxed assumptions; the former assumes events are unordered, but they are competing with each other; the latter assumes events are ordered, i.e., a subject cannot be at risk for event 2 until event 1 occurs. All three models are available for the typical Cox PH model.

As found in the both studies, the lasso Cox regression had some nice features; it provided regularization (variable selection) and shrinkage simultaneously, however it also had some problems, in cases when the number of factors were bigger than the number of observations (n), the lasso Cox regression selects at most n factors; if high correlations existed among factors, the lasso regression only randomly selected one from the correlated group of factors and the prediction performance of the lasso was dominated by ridge regression. Another disadvantage was persistent to all penalized Cox models; the parameter estimates were biased, which would have been even worse if the true unknown parameter was large. As found from the real world case study, polynomial transformations did not improve the model performance for lasso Cox models; inclusion of pair-wise interactions was able to achieve slight improvement to the prediction AUCs for lasso Cox regression.

In general, ridge Cox regression model had very good prediction performance, but it did not select predictors; instead, it shrank most of the coefficients towards zero, but never reached zero, the exceptional prediction performance was achieved by keeping most if not all covariates. Additionally, it was found in the case study that, nonlinear (polynomial) transformation and inclusion of pair-wise interactions did not improve the prediction performance, instead it cost substantial deterioration to the prediction performance of ridge Cox regression.

On the other hand, Elastic-net Cox regression served the purpose for variable selection with relatively good prediction accuracy comparing to lasso and ridge Cox regression. In the real world case study, elastic-net Cox regression models should have relative excellent prediction performance compared to most of the other survival models; however, it was found that the elastic-net interaction model did not perform as good due to the lasso-like penalization terms ($\alpha \approx 1$), which should be considered as an exception.

Otherwise, the elastic-net Cox model was the most effective of all intended survival models.

For the 3 penalized Cox regression models, unfortunately none of them were able to deal with recurrent-event survival data or interval censored survival data; different observations from the same subject was assumed to be independent subjects.

For PCR and PLS Cox model, all covariates were needed for construction of the latent components; unless the model had selected only one or two components, it could be very difficult to distinguish the important predictors from irrelevant covariates, although not all factors were equally important and correlated to the survival outcome. One options to narrow down the relatively more important factors was to apply a soft threshold over the loadings of the factors, however it was a trade-off between prognostic factor detection and the prediction performance; if the threshold was set very high, then only a few important factors could be selected, but the prediction performance would be worse; if a threshold was set too low, the prediction performance could be very good, but many more factors could be left in the model. In addition, the coefficients obtained from the models were corresponding to the constructed components, which was an intermediate between the survival outcome and the original factors; the actual coefficients of the original covariates could be obtained by transforming the coefficients of the latent component with the loading matrix from the models, which made it very difficult to link the survival outcome with the original factors; therefore it was not very intuitive for interpretations.

In the real world case study, the principal components constructed from the original 76 factor was barely estimable with the Cox model, thus the interaction and polynomial transformations were not assessed for PCR analysis. Similar to PLS Cox model, this model was also capable of evaluating unseen variables (the principal components), however because the orthogonal components were constructed independent of the survival outcome, there was no guarantee that the most important factors were selected; if there were too many components needed, the model would become non-estimable. For the simulation study, the prediction performance of the PCR was above average, but it was exceptional for the real world case study since the same dataset was used for both training and testing. Overall, PCR model is not as flexible as most of the survival models,

the results are not intuitively interpretable; the model is useful for variable reduction or clustering, but model performance does not seemed to have much advantage over the other intended models, thus the model is not recommended unless there is no other choice available.

The prediction performance for PLS Cox models was very unsatisfying. However, PLS Cox models were capable of analyzing the uncollected components based on collected factors. Moreover, the model can deal with recurrent-event or competing-risk survival data, since the typical Cox regression model is involved to estimate the coefficients of the PLS components, thus this model should share some benefits from the typical Cox regression analysis: the extensions of Cox model, such as AG, WLW and PWP are also available for PLS Cox model. However, considering the disappointing prediction performance, this model is not recommended unless the intension was to study the unseen variables.

Additionally, two random survival forest (RSF) models, the log-rank based RSF and conditional inference based RSF, were evaluated in the two studies. Both approaches are completely nonparametric, thus they have no model assumptions and should be very flexible with most survival data. Comparing the two RSF models, the conditional inference based RSF model should have slightly better prediction performance for highly correlated survival data; thus the conditional inference based RSF model should be more preferred, even though the log-rank based RSF model could also achieve satisfactory prediction performance.

Chapter 6. Future Work to be Done

In the simulation study, a time-varying treatment effect was noted for the Cox model with *FP* transformation; to resolve the non-proportionality, an interaction between treatment and transformed treatment duration was incorporated, which resolved the non-proportionality for almost all of the factors; but looking at the Martingale residual plot, it was noticed that the distribution of the residuals still showed some patterns before year-2, it was suspected that the transformation for the treatment duration might not have captured all the time-dependent effect, future work will be focused on improvement of the time-varying treatment effect for the model, and further evaluate the impact of the improvement in the time-varying treatment on the model prediction performance, including prediction errors and prediction AUCs.

In the simulation study, some placebo treated subjects had switched treatment during the study; the placebo treated subjects who had switched treatment were considered as two different observations, since the same subject had different treatment over different study period. For this particular case, the switching was independent of the failure event, therefore the AG extension of the Cox PH model was employed; however, in clinical setting, subjects may switch treatment due to lack of efficacy, then the other models such as WLW marginal Cox regression and PWP conditional Cox regression model should have better fit, if subjects had multiple observations or multiple correlated events. Further example of multiple events survival data should be a good topic for evaluation the prediction performance of the three different models.

In the simulation study, the time-varying treatment effect was adjusted using multiple observations per subject; unfortunately the LR- RSF model could not model multiple observations per subject; the CINF-RSF model was able to deal with the situation by assuming that the treatment switching was independent of the failure event, however in clinical studies, time-varying effect may be linked to the outcome, in which case, the CINF-RSF model will not be appropriate. Thus, new RSF models for multiple-event survival data should be very useful.

In both studies, the elastic-net Cox regression models had consistently better prediction performance than most of the other survival models. Unfortunately, the model

cannot handle multiple-event survival data either. It should be more useful, if it can handle multiple event survival data.

In addition to the above approaches, support vector machines was reported to be very efficient for analysis of continuous and categorical outcomes from high dimensional data; thus the learning algorithm on top of Cox regression models may be another effective tool for survival data.

APPENDICES

APPENDIX 1. Polynomial Covariate Terms for Penalized Cox Models – Simulation Study

Total Covariate Terms: 345

Age^4	Age^3	Age^2	Age
Age^2:MAP	Age:MAP	MAP	Age^5
Age:MAP^2	MAP^2	Age^4:MAP	Age^3:MAP
Age:MAP^3	MAP^3	Age^3:MAP^2	Age^2:MAP^2
MAP^5	Age:MAP^4	MAP^4	Age^2:MAP^3
Age^3:BMI	Age^2:BMI	Age:BMI	BMI
Age^2:MAP:BMI	Age:MAP:BMI	MAP:BMI	Age^4:BMI
Age^2:MAP^2:BMI	Age:MAP^2:BMI	MAP^2:BMI	Age^3:MAP:BMI
BMI^2	MAP^4:BMI	Age:MAP^3:BMI	MAP^3:BMI
MAP:BMI^2	Age^3:BMI^2	Age^2:BMI^2	Age:BMI^2
Age:MAP^2:BMI^2	MAP^2:BMI^2	Age^2:MAP:BMI^2	Age:MAP:BMI^2
Age^2:BMI^3	Age:BMI^3	BMI^3	MAP^3:BMI^2
BMI^4	MAP^2:BMI^3	Age:MAP:BMI^3	MAP:BMI^3
Treatment.Active	BMI^5	MAP:BMI^4	Age:BMI^4
Race.White	Race.Hispanic	Race.Asian	Sex.Male
Age:Race.Hispanic	Age:Race.Asian	Age:Sex.Male	Age:Treatment.Active
Age^2:Race.Asian	Age^2:Sex.Male	Age^2:Treatment.Active	Age:Race.White
Age^3:Sex.Male	Age^3:Treatment.Active	Age^2:Race.White	Age^2:Race.Hispanic
Age^4:Treatment.Active	Age^3:Race.White	Age^3:Race.Hispanic	Age^3:Race.Asian
Age^4:Race.White	Age^4:Race.Hispanic	Age^4:Race.Asian	Age^4:Sex.Male
Age^5:Race.Hispanic	Age^5:Race.Asian	Age^5:Sex.Male	Age^5:Treatment.Active
MAP:Race.Asian	MAP:Sex.Male	MAP:Treatment.Active	Age^5:Race.White
Age:MAP:Sex.Male	Age:MAP:Treatment.Active	MAP:Race.White	MAP:Race.Hispanic
Age^2:MAP:Treatment.Active	Age:MAP:Race.White	Age:MAP:Race.Hispanic	Age:MAP:Race.Asian
Age^2:MAP:Race.White	Age^2:MAP:Race.Hispanic	Age^2:MAP:Race.Asian	Age^2:MAP:Sex.Male
Age^3:MAP:Race.Hispanic	Age^3:MAP:Race.Asian	Age^3:MAP:Sex.Male	Age^3:MAP:Treatment.Active
Age^4:MAP:Race.Asian	Age^4:MAP:Sex.Male	Age^4:MAP:Treatment.Active	Age^3:MAP:Race.White
MAP^2:Sex.Male	MAP^2:Treatment.Active	Age^4:MAP:Race.White	Age^4:MAP:Race.Hispanic
Age:MAP^2:Treatment.Active	MAP^2:Race.White	MAP^2:Race.Hispanic	MAP^2:Race.Asian
Age:MAP^2:Race.White	Age:MAP^2:Race.Hispanic	Age:MAP^2:Race.Asian	Age:MAP^2:Sex.Male
Age^2:MAP^2:Race.Hispanic	Age^2:MAP^2:Race.Asian	Age^2:MAP^2:Sex.Male	Age^2:MAP^2:Treatment.Active
Age^3:MAP^2:Race.Asian	Age^3:MAP^2:Sex.Male	Age^3:MAP^2:Treatment.Active	Age^2:MAP^2:Race.White
MAP^3:Sex.Male	MAP^3:Treatment.Active	Age^3:MAP^2:Race.White	Age^3:MAP^2:Race.Hispanic
Age:MAP^3:Treatment.Active	MAP^3:Race.White	MAP^3:Race.Hispanic	MAP^3:Race.Asian
Age:MAP^3:Race.White	Age:MAP^3:Race.Hispanic	Age:MAP^3:Race.Asian	Age:MAP^3:Sex.Male
Age^2:MAP^3:Race.Hispanic	Age^2:MAP^3:Race.Asian	Age^2:MAP^3:Sex.Male	Age^2:MAP^3:Treatment.Active

Appendix 1: Polynomial Covariate Terms for Penalized Cox Models – Simulation Study

Total Covariate Terms: 345

MAP^4:Race.Asian	MAP^4:Sex.Male	MAP^4:Treatment.Active	Age^2:MAP^3:Race.White
Age:MAP^4:Sex.Male	Age:MAP^4:Treatment.Active	MAP^4:Race.White	MAP^4:Race.Hispanic
MAP^5:Treatment.Active	Age:MAP^4:Race.White	Age:MAP^4:Race.Hispanic	Age:MAP^4:Race.Asian
MAP^5:Race.White	MAP^5:Race.Hispanic	MAP^5:Race.Asian	MAP^5:Sex.Male
BMI:Race.Hispanic	BMI:Race.Asian	BMI:Sex.Male	BMI:Treatment.Active
Age:BMI:Race.Asian	Age:BMI:Sex.Male	Age:BMI:Treatment.Active	BMI:Race.White
Age^2:BMI:Sex.Male	Age^2:BMI:Treatment.Active	Age:BMI:Race.White	Age:BMI:Race.Hispanic
Age^3:BMI:Treatment.Active	Age^2:BMI:Race.White	Age^2:BMI:Race.Hispanic	Age^2:BMI:Race.Asian
Age^3:BMI:Race.White	Age^3:BMI:Race.Hispanic	Age^3:BMI:Race.Asian	Age^3:BMI:Sex.Male
Age^4:BMI:Race.Hispanic	Age^4:BMI:Race.Asian	Age^4:BMI:Sex.Male	Age^4:BMI:Treatment.Active
MAP:BMI:Race.Asian	MAP:BMI:Sex.Male	MAP:BMI:Treatment.Active	Age^4:BMI:Race.White
Age:MAP:BMI:Sex.Male	Age:MAP:BMI:Treatment.Active	MAP:BMI:Race.White	MAP:BMI:Race.Hispanic
Age^2:MAP:BMI:Treatment.Active	Age:MAP:BMI:Race.White	Age:MAP:BMI:Race.Hispanic	Age:MAP:BMI:Race.Asian
Age^2:MAP:BMI:Race.White	Age^2:MAP:BMI:Race.Hispanic	Age^2:MAP:BMI:Race.Asian	Age^2:MAP:BMI:Sex.Male
Age^3:MAP:BMI:Race.Hispanic	Age^3:MAP:BMI:Race.Asian	Age^3:MAP:BMI:Sex.Male	Age^3:MAP:BMI:Treatment.Active
MAP^2:BMI:Race.Asian	MAP^2:BMI:Sex.Male	MAP^2:BMI:Treatment.Active	Age^3:MAP:BMI:Race.White
Age:MAP^2:BMI:Sex.Male	Age:MAP^2:BMI:Treatment.Active	MAP^2:BMI:Race.White	MAP^2:BMI:Race.Hispanic
Age^2:MAP^2:BMI:Treatment.Active	Age:MAP^2:BMI:Race.White	Age:MAP^2:BMI:Race.Hispanic	Age:MAP^2:BMI:Race.Asian
Age^2:MAP^2:BMI:Race.White	Age^2:MAP^2:BMI:Race.Hispanic	Age^2:MAP^2:BMI:Race.Asian	Age^2:MAP^2:BMI:Sex.Male
MAP^3:BMI:Race.Hispanic	MAP^3:BMI:Race.Asian	MAP^3:BMI:Sex.Male	MAP^3:BMI:Treatment.Active
Age:MAP^3:BMI:Race.Asian	Age:MAP^3:BMI:Sex.Male	Age:MAP^3:BMI:Treatment.Active	MAP^3:BMI:Race.White
MAP^4:BMI:Sex.Male	MAP^4:BMI:Treatment.Active	Age:MAP^3:BMI:Race.White	Age:MAP^3:BMI:Race.Hispanic
BMI^2:Treatment.Active	MAP^4:BMI:Race.White	MAP^4:BMI:Race.Hispanic	MAP^4:BMI:Race.Asian
BMI^2:Race.White	BMI^2:Race.Hispanic	BMI^2:Race.Asian	BMI^2:Sex.Male
Age:BMI^2:Race.Hispanic	Age:BMI^2:Race.Asian	Age:BMI^2:Sex.Male	Age:BMI^2:Treatment.Active
Age^2:BMI^2:Race.Asian	Age^2:BMI^2:Sex.Male	Age^2:BMI^2:Treatment.Active	Age:BMI^2:Race.White
Age^3:BMI^2:Sex.Male	Age^3:BMI^2:Treatment.Active	Age^2:BMI^2:Race.White	Age^2:BMI^2:Race.Hispanic
MAP:BMI^2:Treatment.Active	Age^3:BMI^2:Race.White	Age^3:BMI^2:Race.Hispanic	Age^3:BMI^2:Race.Asian
MAP:BMI^2:Race.White	MAP:BMI^2:Race.Hispanic	MAP:BMI^2:Race.Asian	MAP:BMI^2:Sex.Male
Age:MAP:BMI^2:Race.Hispanic	Age:MAP:BMI^2:Race.Asian	Age:MAP:BMI^2:Sex.Male	Age:MAP:BMI^2:Treatment.Active
Age^2:MAP:BMI^2:Race.Asian	Age^2:MAP:BMI^2:Sex.Male	Age^2:MAP:BMI^2:Treatment.Active	Age:MAP:BMI^2:Race.White
MAP^2:BMI^2:Sex.Male	MAP^2:BMI^2:Treatment.Active	Age^2:MAP:BMI^2:Race.White	Age^2:MAP:BMI^2:Race.Hispanic
Age:MAP^2:BMI^2:Treatment.Active	MAP^2:BMI^2:Race.White	Age^2:MAP:BMI^2:Race.Hispanic	MAP^2:BMI^2:Race.Asian
Age:MAP^2:BMI^2:Race.White	Age:MAP^2:BMI^2:Race.Hispanic	Age:MAP^2:BMI^2:Race.Asian	Age:MAP^2:BMI^2:Sex.Male
Age^2:MAP^2:BMI^2:Race.Hispanic	Age^2:MAP^2:BMI^2:Race.Asian	MAP^3:BMI^2:Sex.Male	MAP^3:BMI^2:Treatment.Active
MAP^3:BMI^2:Race.Hispanic	MAP^3:BMI^2:Race.Asian	BMI^3:Treatment.Active	MAP^3:BMI^2:Race.White
BMI^3:Race.Asian	BMI^3:Sex.Male		

Appendix 1: Polynomial Covariate Terms for Penalized Cox Models – Simulation Study

Total Covariate Terms: 345

Age:BMI^3:Sex.Male	Age:BMI^3:Treatment.Active	BMI^3:Race.White	BMI^3:Race.Hispanic
Age^2:BMI^3:Treatment.Active	Age:BMI^3:Race.White	Age:BMI^3:Race.Hispanic	Age:BMI^3:Race.Asian
Age^2:BMI^3:Race.White	Age^2:BMI^3:Race.Hispanic	Age^2:BMI^3:Race.Asian	Age^2:BMI^3:Sex.Male
MAP:BMI^3:Race.Hispanic	MAP:BMI^3:Race.Asian	MAP:BMI^3:Sex.Male	MAP:BMI^3:Treatment.Active
Age:MAP:BMI^3:Race.Asian	Age:MAP:BMI^3:Sex.Male	Age:MAP:BMI^3:Treatment.Active	MAP:BMI^3:Race.White
MAP^2:BMI^3:Sex.Male	MAP^2:BMI^3:Treatment.Active	Age:MAP:BMI^3:Race.White	Age:MAP:BMI^3:Race.Hispanic
BMI^4:Treatment.Active	MAP^2:BMI^3:Race.White	MAP^2:BMI^3:Race.Hispanic	MAP^2:BMI^3:Race.Asian
BMI^4:Race.White	BMI^4:Race.Hispanic	BMI^4:Race.Asian	BMI^4:Sex.Male
Age:BMI^4:Race.Hispanic	Age:BMI^4:Race.Asian	Age:BMI^4:Sex.Male	Age:BMI^4:Treatment.Active
MAP:BMI^4:Race.Asian	MAP:BMI^4:Sex.Male	MAP:BMI^4:Treatment.Active	Age:BMI^4:Race.White
BMI^5:Sex.Male	BMI^5:Treatment.Active	MAP:BMI^4:Race.White	MAP:BMI^4:Race.Hispanic
Treatment.Active:Sex.Male	BMI^5:Race.White	BMI^5:Race.Hispanic	BMI^5:Race.Asian
Sex.Male:Race.Asian	Treatment.Active:Race.White	Treatment.Active:Race.Hispanic	Treatment.Active:Race.Asian
Race.Asian:Race.White	Race.Asian:Race.Hispanic	Sex.Male:Race.White	Sex.Male:Race.Hispanic

APPENDIX 2. Covariate Terms for the "Best" Lasso Cox Regression Model Cross Validated via Partial Log Likelihood Deviance – Simulation Study

Terms	MAP^4	Age:MAP^2	MAP	Age^2
Coefficients	0.0041	0.0042	0.1092	0.005
Terms	Treatment=Active	Age:MAP:BMI^2	Age^3:MAP:BMI	MAP^5
Coefficients	-0.3509	0.0254	0.0005	0.0025
Terms	MAP^5:Sex=Male	Race=White	Race=Asian	Sex=Male
Coefficients	0.0014	-0.0513	-0.0084	-0.135
Terms	Age^3:MAP:BMI:Sex=Male	Age:MAP:BMI:Sex=Male	Age^2:BMI:Race=Asian	BMI:Race=White
Coefficients	0.0007	0.0214	-0.0169	-0.0021
Terms	BMI^2:Race=Hispanic	BMI^2:Sex=Male	MAP^3:BMI:Sex=Male	Age^3:MAP:BMI:Race=Hispanic
Coefficients	-0.0045	-0.0264	0.0011	0.0118
Terms	Sex.Male:Race=White	Treatment.Active:Race=His	Age^2:BMI^2:Race=Asian	
Coefficients	-0.0333	-0.0329	-0.02	

Which can be formulated in Cox regression model as $\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}$, where $X\hat{\beta} =$ $0.005\ \text{Age}^2 + 0.1092\ \text{MAP} + 0.0042\ \text{Age: MAP}^2 + 0.0041\ \text{MAP}^4 + 0.0025\ \text{MAP}^5 + 5\text{e} -$ $04\ \text{Age}^3\text{: MAP: BMI} + 0.0254\ \text{Age: MAP: BMI}^2 + -0.3509\ \text{Treatment (Active)} + -0.135\ \text{Sex (Male)} +$ $-0.0084\ \text{Race (Asian)} - 0.0513\ \text{Race (White)} + 0.0014\ \text{MAP}^5\text{: Sex(Male)} +$ $-0.0021\ \text{BMI: Race(White)} - 0.0169\ \text{Age}^2\text{: BMI: Race (Asian)} + 0.0214\ \text{Age: MAP: BMI: Sex(Male)} +$ $7\text{e} - 04\ \text{Age}^3\text{: MAP: BMI: Sex(Male)} + 0.0118\ \text{Age}^3\text{: MAP: BMI: Race (Hispanic)} +$ $0.0011\ \text{MAP}^3\text{: BMI: Sex(Male)} - 0.0264\ \text{BMI}^2\text{: Sex (Male)} - 0.0045\ \text{BMI}^2\text{: Race (Hispanic)} -$ $0.02\ \text{Age}^2\text{: BMI}^2\text{: Race (Asian)} - 0.0329\ \text{Treatment (Active): Race (Hispanic)} -$ $0.0333\ \text{Sex (Male): Race (White)}$

APPENDIX 3. Covariate Terms for the "Best" Ridge Cox Regression Model Cross Validated via Partial Log Likelihood Deviance – Simulation Study

Total Covariate Terms: 342

Terms	Age^4	Age^3	Age^2	Age
Coefficients	0.0008	0.0007	0.0069	0.0053
Terms	Age^2:MAP	Age:MAP	MAP	Age^5
Coefficients	0.0046	0.0049	0.0206	0
Terms	Age:MAP^2	MAP^2	Age^4:MAP	Age^3:MAP
Coefficients	0.0029	0.003	0.0001	0.0004
Terms	Age:MAP^3	MAP^3	Age^3:MAP^2	Age^2:MAP^2
Coefficients	0.0014	0.0044	0.0004	0.0016
Terms	MAP^5	Age:MAP^4	MAP^4	Age^2:MAP^3
Coefficients	0.0007	0.0006	0.0015	0.0011
Terms	Age^3:BMI	Age^2:BMI	Age:BMI	BMI
Coefficients	-0.0017	0.0004	0.0041	0.0024
Terms	Age^2:MAP:BMI	Age:MAP:BMI	MAP:BMI	Age^4:BMI
Coefficients	0.0027	0.0033	-0.0011	0.0004
Terms	Age^2:MAP^2:BMI	Age:MAP^2:BMI	MAP^2:BMI	Age^3:MAP:BMI
Coefficients	0.0007	0.004	-0.0001	0.002
Terms	BMI^2	MAP^4:BMI	Age:MAP^3:BMI	MAP^3:BMI
Coefficients	-0.0065	0.0002	0.001	0.0022
Terms	MAP:BMI^2	Age^3:BMI^2	Age^2:BMI^2	Age:BMI^2
Coefficients	0.0048	0.0007	0.0021	0.0001
Terms	Age:MAP^2:BMI^2	MAP^2:BMI^2	Age^2:MAP:BMI^2	Age:MAP:BMI^2
Coefficients	0.0016	-0.0013	0.0012	0.006
Terms	Age^2:BMI^3	Age:BMI^3	BMI^3	MAP^3:BMI^2
Coefficients	0.0001	0.001	-0.0005	0.0007
Terms	BMI^4	MAP^2:BMI^3	Age:MAP:BMI^3	MAP:BMI^3
Coefficients	-0.0001	-0.0002	0.0007	-0.0002
Terms	Treatment=Active	BMI^5	MAP:BMI^4	Age:BMI^4
Coefficients	-0.0699	-0.0001	0.0003	0
Terms	Race=White	Race=Hispanic	Race=Asian	Sex=Male
Coefficients	-0.0271	-0.0155	-0.0319	-0.0432
Terms	Age:Race=Hispanic	Age:Race=Asian	Age:Sex=Male	Age:Treatment=Active
Coefficients	0.0184	-0.0201	0.0043	0.0017
Terms	Age^2:Race=Asian	Age^2:Sex=Male	Age^2:Treatment=Active	Age:Race=White
Coefficients	-0.0155	0.0007	-0.0077	-0.0007
Terms	Age^3:Sex=Male	Age^3:Treatment=Active	Age^2:Race=White	Age^2:Race=Hispanic
Coefficients	-0.0002	-0.0011	0.0006	0.0034

Appendix 3: Covariate Terms for the "Best" Ridge Cox Regression Model Cross Validated via Partial Log Likelihood Deviance – Simulation Study

Total Covariate Terms: 342

Terms	Age^4:Treatment=Active	Age^3:Race=White	Age^3:Race=Hispanic	Age^3:Race=Asian
Coefficients	-0.0002	-0.0014	0.0044	-0.0037
Terms	Age^4:Race=White	Age^4:Race=Hispanic	Age^4:Race=Asian	Age^4:Sex=Male
Coefficients	0.0002	0.0018	-0.0027	0.0001
Terms	Age^5:Race=Hispanic	Age^5:Race=Asian	Age^5:Sex=Male	Age^5:Treatment=Active
Coefficients	0.0006	-0.0008	-0.0002	-0.0003
Terms	MAP:Race=Asian	MAP:Sex=Male	MAP:Treatment=Active	Age^5:Race=White
Coefficients	-0.007	0.0125	0.0034	-0.0001
Terms	Age:MAP:Sex=Male	Age:MAP:Treatment=Active	MAP:Race=White	MAP:Race=Hispanic
Coefficients	0.0016	0.0006	0.0187	0.0168
Terms	Age^2:MAP:Treatment=Active	Age:MAP:Race=White	Age:MAP:Race=Hispanic	Age:MAP:Race=Asian
Coefficients	-0.0012	0.0028	-0.0195	-0.0139
Terms	Age^2:MAP:Race=White	Age^2:MAP:Race=Hispanic	Age^2:MAP:Race=Asian	Age^2:MAP:Sex=Male
Coefficients	0.0005	0.0071	-0.0124	0.0005
Terms	Age^3:MAP:Race=Hispanic	Age^3:MAP:Race=Asian	Age^3:MAP:Sex=Male	Age^3:MAP:Treatment=Active
Coefficients	-0.0026	0.0016	0.0002	-0.0021
Terms	Age^4:MAP:Race=Asian	Age^4:MAP:Sex=Male	Age^4:MAP:Treatment=Active	Age^3:MAP:Race=White
Coefficients	-0.0021	-0.0005	-0.0004	-0.0021
Terms	MAP^2:Sex=Male	MAP^2:Treatment=Active	Age^4:MAP:Race=White	Age^4:MAP:Race=Hispanic
Coefficients	-0.0026	-0.0084	-0.0013	0.0009
Terms	Age:MAP^2:Treatment=Active	MAP^2:Race=White	MAP^2:Race=Hispanic	MAP^2:Race=Asian
Coefficients	-0.0002	-0.0031	-0.0001	-0.001
Terms	Age:MAP^2:Race=White	Age:MAP^2:Race=Hispanic	Age:MAP^2:Race=Asian	Age:MAP^2:Sex=Male
Coefficients	0.0003	-0.0015	-0.0014	0.0042
Terms	Age^2:MAP^2:Race=Hispanic	Age^2:MAP^2:Race=Asian	Age^2:MAP^2:Sex=Male	Age^2:MAP^2:Treatment=Active
Coefficients	0.0032	0.007	0.0004	-0.0011
Terms	Age^3:MAP^2:Race=Asian	Age^3:MAP^2:Sex=Male	Age^3:MAP^2:Treatment=Active	Age^2:MAP^2:Race=White
Coefficients	0.0005	0.0004	-0.0004	-0.0023
Terms	MAP^3:Sex=Male	MAP^3:Treatment=Active	Age^3:MAP^2:Race=White	Age^3:MAP^2:Race=Hispanic
Coefficients	0.004	0.0007	-0.0018	0.0005
Terms	Age:MAP^3:Treatment=Active	MAP^3:Race=White	MAP^3:Race=Hispanic	MAP^3:Race=Asian
Coefficients	0.0007	0.0043	0.0036	0.0027
Terms	Age:MAP^3:Race=White	Age:MAP^3:Race=Hispanic	Age:MAP^3:Race=Asian	Age:MAP^3:Sex=Male
Coefficients	0.0003	-0.0009	-0.006	0.0006
Terms	Age^2:MAP^3:Race=Hispanic	Age^2:MAP^3:Race=Asian	Age^2:MAP^3:Sex=Male	Age^2:MAP^3:Treatment=Active
Coefficients	0.0011	-0.0021	0.001	0.0003

Appendix 3: Covariate Terms for the "Best" Ridge Cox Regression Model Cross Validated via Partial Log Likelihood Deviance – Simulation Study

Total Covariate Terms: 342

Terms	MAP^4:Race=Asian	MAP^4:Sex=Male	MAP^4:Treatment=Active	Age^2:MAP^3:Race=White
Coefficients	-0.0006	0.0013	0.0006	0.0007
Terms	Age:MAP^4:Sex=Male	Age:MAP^4:Treatment=Acti	MAP^4:Race=White	MAP^4:Race=Hispanic
Coefficients	0.0008	-0.0001	0.001	0.0016
Terms	MAP^5:Treatment=Active	Age:MAP^4:Race=White	Age:MAP^4:Race=Hispanic	Age:MAP^4:Race=Asian
Coefficients	0.0002	-0.0003	0.0001	0.0007
Terms	MAP^5:Race=White	MAP^5:Race=Hispanic	MAP^5:Race=Asian	MAP^5:Sex=Male
Coefficients	0.0007	0.0008	0.0007	0.0007
Terms	BMI:Race=Hispanic	BMI:Race=Asian	BMI:Sex=Male	BMI:Treatment=Active
Coefficients	0.0067	0.0051	0.0073	0.0007
Terms	Age:BMI:Race=Asian	Age:BMI:Sex=Male	Age:BMI:Treatment=Active	BMI:Race=White
Coefficients	0.0062	0.0056	0.0041	-0.0093
Terms	Age^2:BMI:Sex=Male	Age^2:BMI:Treatment=Acti	Age:BMI:Race=White	Age:BMI:Race=Hispanic
Coefficients	0.0004	-0.0022	0.0001	0.0085
Terms	Age^3:BMI:Treatment=Active	Age^2:BMI:Race=White	Age^2:BMI:Race=Hispanic	Age^2:BMI:Race=Asian
Coefficients	-0.0018	-0.0038	-0.0015	-0.0248
Terms	Age^3:BMI:Race=White	Age^3:BMI:Race=Hispanic	Age^3:BMI:Race=Asian	Age^3:BMI:Sex=Male
Coefficients	-0.0028	0.0002	-0.0018	-0.0028
Terms	Age^4:BMI:Race=Hispanic	Age^4:BMI:Race=Asian	Age^4:BMI:Sex=Male	Age^4:BMI:Treatment=Active
Coefficients	0.0002	-0.0068	0.0004	-0.0002
Terms	MAP:BMI:Race=Asian	MAP:BMI:Sex=Male	MAP:BMI:Treatment=Active	Age^4:BMI:Race=White
Coefficients	0.0201	-0.0021	-0.0029	-0.0003
Terms	Age:MAP:BMI:Sex=Male	Age:MAP:BMI:Treatment=A	MAP:BMI:Race=White	MAP:BMI:Race=Hispanic
Coefficients	0.0087	-0.005	0.0018	0.0055
Terms	Age^2:MAP:BMI:Treatment=Active	Age:MAP:BMI:Race=White	Age:MAP:BMI:Race=Hispanic	Age:MAP:BMI:Race=Asian
Coefficients	0.0038	-0.0002	0.0128	-0.0042
Terms	Age^2:MAP:BMI:Race=White	Age^2:MAP:BMI:Race=Hisp	Age^2:MAP:BMI:Race=Asian	Age^2:MAP:BMI:Sex=Male
Coefficients	0.0017	0.0097	0.0061	0.0001
Terms	Age^3:MAP:BMI:Race=Hispanic	Age^3:MAP:BMI:Race=Asia	Age^3:MAP:BMI:Sex=Male	Age^3:MAP:BMI:Treatment=A
Coefficients	0.0053	-0.0026	0.0027	-0.0001
Terms	MAP^2:BMI:Race=Asian	MAP^2:BMI:Sex=Male	MAP^2:BMI:Treatment=Active	Age^3:MAP:BMI:Race=White
Coefficients	-0.0037	-0.0022	-0.0012	0.0009
Terms	Age:MAP^2:BMI:Sex=Male	Age:MAP^2:BMI:Treatment	MAP^2:BMI:Race=White	MAP^2:BMI:Race=Hispanic
Coefficients	0.0048	-0.0017	-0.0041	0.0054
Terms	Age^2:MAP^2:BMI:Treatment=Acti	Age:MAP^2:BMI:Race=Whit	Age:MAP^2:BMI:Race=Hispanic	Age:MAP^2:BMI:Race=Asian
Coefficients	0.0001	0.0002	0.0054	0.0122

Appendix 3: Covariate Terms for the "Best" Ridge Cox Regression Model Cross Validated via Partial Log Likelihood Deviance – Simulation Study

Total Covariate Terms: 342

Terms	Age^2:MAP^2:BMI:Race=White	Age^2:MAP^2:BMI:Race=Hispanic	Age^2:MAP^2:BMI:Race=Asian	Age^2:MAP^2:BMI:Sex=Male	
Coefficients	-0.0017	-0.002	-0.0047	0.0014	
Terms	MAP^3:BMI:Race=Hispanic	MAP^3:BMI:Race=Asian	MAP^3:BMI:Sex=Male	MAP^3:BMI:Treatment=Active	
Coefficients	0.0008	0.0038	0.0029	0.0023	
Terms	Age:MAP^3:BMI:Race=Asian	Age:MAP^3:BMI:Sex=Male	Age:MAP^3:BMI:Treatment=Ac	MAP^3:BMI:Race=White	
Coefficients	-0.0016	0.0011	-0.0017	0.0027	
Terms	MAP^4:BMI:Sex=Male	MAP^4:BMI:Treatment=Active	Age:MAP^3:BMI:Race=White	Age:MAP^3:BMI:Race=Hispanic	
Coefficients	0.0001	0.0001	-0.0009	0.0025	
Terms	BMI^2:Treatment=Active	MAP^4:BMI:Race=White	MAP^4:BMI:Race=Hispanic	MAP^4:BMI:Race=Asian	
Coefficients	-0.0122	-0.0007	0.002	-0.0008	
Terms	BMI^2:Race=White	BMI^2:Race=Hispanic	BMI^2:Race=Asian	BMI^2:Sex=Male	
Coefficients	-0.0036	-0.0121	-0.0116	-0.0111	
Terms	Age:BMI^2:Race=Hispanic	Age:BMI^2:Race=Asian	Age:BMI^2:Sex=Male	Age:BMI^2:Treatment=Active	
Coefficients	-0.0033	-0.0026	0.0023	0.0003	
Terms	Age^2:BMI^2:Race=Asian	Age^2:BMI^2:Sex=Male	Age^2:BMI^2:Treatment=Activ	Age:BMI^2:Race=White	
Coefficients	-0.0142	-0.0008	0.0017	0.002	
Terms	Age^3:BMI^2:Sex=Male	Age^3:BMI^2:Treatment=Active	Age^2:BMI^2:Race=White	Age^2:BMI^2:Race=Hispanic	
Coefficients	0.0011	0.0004	0.0029	0.0012	
Terms	MAP:BMI^2:Treatment=Active	Age^3:BMI^2:Race=White	Age^3:BMI^2:Race=Hispanic	Age^3:BMI^2:Race=Asian	
Coefficients	0.0034	0.0002	0.0012	0.0019	
Terms	MAP:BMI^2:Race=White	MAP:BMI^2:Race=Hispanic	MAP:BMI^2:Race=Asian	MAP:BMI^2:Sex=Male	
Coefficients	0.0003	0.0099	0.0012	0.0021	
Terms	Age:MAP:BMI^2:Race=Hispanic	Age:MAP:BMI^2:Race=Asian	Age:MAP:BMI^2:Sex=Male	Age:MAP:BMI^2:Treatment=Active	
Coefficients	0.0022	0.0024	0.0077	0.0016	
Terms	Age^2:MAP:BMI^2:Race=Asian	Age^2:MAP:BMI^2:Sex=Male	Age^2:MAP:BMI^2:Treatment=	Age:MAP:BMI^2:Race=White	
Coefficients	-0.0079	-0.0006	0.0002	0.0054	
Terms	MAP^2:BMI^2:Sex=Male	MAP^2:BMI^2:Treatment=Active	Age^2:MAP:BMI^2:Race=White	Age^2:MAP:BMI^2:Race=Hispanic	
Coefficients	-0.0018	-0.0021	-0.0009	0.0031	
Terms	Age:MAP^2:BMI^2:Treatment=A	MAP^2:BMI^2:Race=White	MAP^2:BMI^2:Race=Hispanic	MAP^2:BMI^2:Race=Asian	
Coefficients	0.0001	-0.0002	-0.0028	0	
Terms	Age:MAP^2:BMI^2:Race=White	Age:MAP^2:BMI^2:Race=Hispanic	Age:MAP^2:BMI^2:Race=Asian	Age:MAP^2:BMI^2:Sex=Male	
Coefficients	0.0017	-0.0032	0.0036	0.0015	
Terms	MAP^3:BMI^2:Race=Hispanic	MAP^3:BMI^2:Race=Asian	MAP^3:BMI^2:Sex=Male	MAP^3:BMI^2:Treatment=Active	
Coefficients	0.002	0.0003	0.0007	0.0004	
Terms	BMI^3:Race=Asian	BMI^3:Sex=Male	BMI^3:Treatment=Active	MAP^3:BMI^2:Race=White	
Coefficients	0.001	-0.0002	-0.0003	-0.0001	

Appendix 3: Covariate Terms for the "Best" Ridge Cox Regression Model Cross Validated via Partial Log Likelihood Deviance – Simulation Study

Total Covariate Terms: 342

Terms	Age:BMI^3:Sex=Male	Age:BMI^3:Treatment=Active	BMI^3:Race=White	BMI^3:Race=Hispanic
Coefficients	0.0001	0.0001	-0.002	-0.0022
Terms	Age^2:BMI^3:Treatment=Active	Age:BMI^3:Race=White	Age:BMI^3:Race=Hispanic	Age:BMI^3:Race=Asian
Coefficients	0	0.001	0	-0.0056
Terms	Age^2:BMI^3:Race=White	Age^2:BMI^3:Race=Hispanic	Age^2:BMI^3:Race=Asian	Age^2:BMI^3:Sex=Male
Coefficients	-0.0001	-0.0014	-0.0053	-0.0003
Terms	MAP:BMI^3:Race=Hispanic	MAP:BMI^3:Race=Asian	MAP:BMI^3:Sex=Male	MAP:BMI^3:Treatment=Active
Coefficients	0.0032	-0.0015	-0.0012	-0.0007
Terms	Age:MAP:BMI^3:Race=Asian	Age:MAP:BMI^3:Sex=Male	Age:MAP:BMI^3:Treatment=Active	MAP:BMI^3:Race=White
Coefficients	0.0019	0.0012	-0.001	0.0013
Terms	MAP^2:BMI^3:Sex=Male	MAP^2:BMI^3:Treatment=Active	Age:MAP:BMI^3:Race=White	Age:MAP:BMI^3:Race=Hispanic
Coefficients	-0.0012	-0.0001	0.0006	0.0047
Terms	BMI^4:Treatment=Active	MAP^2:BMI^3:Race=White	MAP^2:BMI^3:Race=Hispanic	MAP^2:BMI^3:Race=Asian
Coefficients	-0.0001	-0.0004	-0.0016	-0.0009
Terms	BMI^4:Race=White	BMI^4:Race=Hispanic	BMI^4:Race=Asian	BMI^4:Sex=Male
Coefficients	0.0003	0	-0.002	-0.0006
Terms	Age:BMI^4:Race=Hispanic	Age:BMI^4:Race=Asian	Age:BMI^4:Sex=Male	Age:BMI^4:Treatment=Active
Coefficients	0.0001	0.0017	0.0002	0.0001
Terms	MAP:BMI^4:Race=Asian	MAP:BMI^4:Sex=Male	MAP:BMI^4:Treatment=Active	Age:BMI^4:Race=White
Coefficients	-0.0005	-0.0004	0.0005	0.0004
Terms	BMI^5:Sex=Male	BMI^5:Treatment=Active	MAP:BMI^4:Race=White	MAP:BMI^4:Race=Hispanic
Coefficients	-0.0001	0	0	0
Terms	Treatment=Active:Sex=Male	BMI^5:Race=White	BMI^5:Race=Hispanic	BMI^5:Race=Asian
Coefficients	-0.0592	-0.0001	-0.0002	0.0001
Terms	Sex=Male:Race=Asian	Treatment=Active:Race=White	Treatment=Active:Race=Hispanic	Treatment=Active:Race=Asian
Coefficients	-0.0469	-0.0514	-0.0586	-0.0573
Terms	Sex=Male:Race=White	Sex=Male:Race=Hispanic		
Coefficients	-0.0399	-0.0178		

The exact formula of this model can be easily programmed, but they are not presented, because the approach selects too many covariates, it is difficult to fit the entire model in one or two pages.

APPENDIX 4. Covariate Terms for Partial Least Squares Cox Regression Model – Simulation Study

Total Covariate Terms=52

SBP	MAP	BMP	Age
Race=White	Sex=Male	Treatment=Active	DBP
Age:MAP	Age:BMI	Race=Asian	Race=Hispanic
Age:Sex=Male	Age:Treatment=Active	Age:DBP	Age:SBP
BMI:MAP	Age:Race=Asian	Age:Race=Hispanic	Age:Race=White
BMI:Sex=Male	BMI:Treatment=Active	BMI:DBP	BMI:SBP
MAP:SBP	BMI:Race=Asian	BMI:Race=Hispanic	BMI:Race=White
MAP:Race=White	MAP:Sex=Male	MAP:Treatment=Active	MAP:DBP
SBP:Treatment=Active	SBP:DBP	MAP:Race=Asian	MAP:Race=Hispanic
SBP:Race=Asian	SBP:Race=Hispanic	SBP:Race=White	SBP:Sex=Male
DBP:Race=Hispanic	DBP:Race=White	DBP:Sex=Male	DBP:Treatment=Active
Treatment=Active:Race=Hispanic	Treatment=Active:Race=White	Treatment=Active:Sex=Male	DBP:Race=Asian
Sex=Male:Race=Asian	Sex=Male:Race=Hispanic	Sex=Male:Race=White	Treatment=Active:Race=Asian

APPENDIX 5. Regression Coefficients for all Factor in the Original Scale for PLS Cox Model – Simulation Study

Terms:	Coef	SD.Coef	Terms:	Coef	SD.Coef	Terms:	Coef	SD.Coef
BMI:Race=White	-0.005	-0.0679	BMI:DBP	0	-0.013	Age	0	0.0031
MAP:Race=Asian	-0.002	-0.0371	BMI:Treatment=Active	-0.003	-0.0478	BMI	0	0.0003
MAP:Race=Hispanic	-0.001	-0.046	BMI:Sex=Male	-0.002	-0.0271	MAP	0.006	0.0378
MAP:Race=White	0	-0.0204	MAP:SBP	0	0.0705	SBP	0.011	0.0945
SBP:Race=Asian	-0.002	-0.0436	MAP:DBP	0	0.0116	DBP	-0.003	-0.0136
SBP:Race=Hispanic	-0.001	-0.0405	MAP:Treatment=Active	-0.002	-0.0659	Treatment=Active	-0.073	-0.0361
SBP:Race=White	0	-0.0214	MAP:Sex=Male	-0.001	-0.0402	Sex=Male	-0.094	-0.0459
DBP:Race=Asian	-0.002	-0.0322	SBP:DBP	0	0.0473	Race=Asian	0.021	0.0049
DBP:Race=Hispanic	-0.002	-0.0499	SBP:Treatment=Active	-0.001	-0.0616	Race=Hispanic	-0.123	-0.047
DBP:Race=White	-0.001	-0.0197	SBP:Sex=Male	-0.001	-0.0399	Race=White	-0.034	-0.0167
Treatment=Active:Race=Asian	0.084	0.0166	DBP:Treatment=Active	-0.002	-0.069	Age:BMI	0	-0.0106
Treatment=Active:Race=Hispanic	0.064	0.0189	DBP:Sex=Male	-0.001	-0.0403	Age:MAP	0	0.0254
Treatment=Active:Race=White	0.213	0.0975	Treatment=Active:Sex=	0.097	0.0456	Age:SBP	0	0.0406
Sex=Male:Race=Asian	0.056	0.0107	Age:Race=Asian	0.005	0.0517	Age:DBP	0	0.0144
Sex=Male:Race=Hispanic	0.284	0.086	Age:Race=Hispanic	-0.002	-0.0416	Age:Treatment=Active	-0.001	-0.0352
Sex=Male:Race=White	0.009	0.0041	Age:Race=White	-0.005	-0.1299	Age:Sex=Male	-0.001	-0.0329
BMI:Race=Asian	-0.006	-0.0389	BMI:MAP	0	0.0017			
BMI:Race=Hispanic	0	0.0009	BMI:SBP	0	0.0203			

APPENDIX 6. Descriptive Summary of the NKI70 Data

Total Gene Signatures = 70

Rows	Factors	n	Mean ± STD	Median	0.25	0.75	Range
1	TSPYL5	144	-0.1085 ± 0.32979	-0.0890	-0.3308	0.1173	-1.0830, 0.6018
2	Contig63649.RC	144	-0.0539 ± 0.23664	-0.0956	-0.2311	0.0946	-0.5077, 0.7757
3	DIAPH3	144	-0.0325 ± 0.23532	-0.0217	-0.1786	0.1271	-0.6789, 0.6178
4	NUSAP1	144	-0.0564 ± 0.26789	-0.0316	-0.2177	0.1478	-0.7968, 0.5067
5	AA555029.RC	144	-0.0550 ± 0.17105	-0.0582	-0.1702	0.0586	-0.4778, 0.3741
6	ALDH4A1	144	0.0106 ± 0.15575	0.0005	-0.0777	0.0806	-0.3716, 0.4642
7	QSCN6L1	144	-0.0396 ± 0.21149	-0.0692	-0.1927	0.0868	-0.4423, 0.6090
8	FGF18	144	-0.0602 ± 0.26778	-0.0698	-0.2583	0.1152	-0.7682, 0.7049
9	DIAPH3.1	144	-0.0021 ± 0.23399	-0.0093	-0.1607	0.1607	-0.5863, 0.6765
10	Contig32125.RC	144	-0.0221 ± 0.16264	-0.0333	-0.1295	0.0944	-0.4585, 0.5607
11	BBC3	144	0.0011 ± 0.16126	0.0132	-0.1129	0.1127	-0.4510, 0.3669
12	DIAPH3.2	144	0.0078 ± 0.15621	-0.001	-0.0825	0.1033	-0.5636, 0.4791
13	RP5.860F19.3	144	-0.0365 ± 0.20715	-0.0486	-0.157	0.088	-0.4943, 0.5168
14	C16orf61	144	-0.0302 ± 0.17294	-0.0459	-0.1564	0.077	-0.4444, 0.4842
15	SCUBE2	144	-0.2006 ± 0.52490	-0.1451	-0.6436	0.2534	-1.2650, 0.8919
16	EXT1	144	-0.0180 ± 0.13176	-0.0215	-0.0829	0.0574	-0.3552, 0.3306
17	FLT1	144	-0.0125 ± 0.16900	-0.0058	-0.1177	0.0933	-0.4493, 0.3549
18	GNAZ	144	-0.0494 ± 0.22936	-0.0611	-0.1825	0.0744	-0.7898, 0.8189
19	OXCT1	144	-0.0459 ± 0.16211	-0.0338	-0.1525	0.0442	-0.4804, 0.4580
20	MMP9	144	-0.3217 ± 0.27493	-0.3403	-0.5287	-0.1596	-0.9087, 0.5985
21	RUNDC1	144	-0.0197 ± 0.18572	0.0021	-0.16	0.1075	-0.4307, 0.8201
22	Contig35251.RC	144	-0.0368 ± 0.19719	-0.0595	-0.1828	0.0671	-0.4311, 0.5975
23	ECT2	144	0.0083 ± 0.19963	-0.0144	-0.1455	0.1258	-0.3594, 0.5561
24	GMPS	144	-0.0621 ± 0.19410	-0.0945	-0.1911	0.0587	-0.6119, 0.5941
25	KNTC2	144	0.0174 ± 0.19725	0.0087	-0.1072	0.1071	-0.4242, 0.5938
26	WISP1	144	0.0042 ± 0.17785	0.0227	-0.1238	0.1216	-0.4715, 0.4185
27	CDC42BPA	144	-0.0184 ± 0.16753	-0.0129	-0.1368	0.0978	-0.3797, 0.4393
28	SERF1A	144	0.0086 ± 0.15048	-0.0164	-0.0998	0.0801	-0.3175, 0.4306
29	AYTL2	144	-0.0028 ± 0.14121	-0.0096	-0.0911	0.0804	-0.4152, 0.4397
30	GSTM3	144	-0.0845 ± 0.27873	-0.0792	-0.2933	0.0774	-0.6756, 0.6406
31	GPR180	144	-0.0610 ± 0.20079	-0.0625	-0.194	0.062	-0.5987, 0.5602
32	RAB6B	144	-0.0646 ± 0.28632	-0.1454	-0.2613	0.0212	-0.5109, 0.8648
33	ZNF533	144	-0.2450 ± 0.43754	-0.4011	-0.5925	0.0461	-0.9177, 0.9944
34	RTN4RL1	144	0.0160 ± 0.18066	0.0093	-0.0849	0.1234	-0.4278, 0.6491
35	UCHL5	144	-0.0132 ± 0.13884	-0.0311	-0.1009	0.0889	-0.3654, 0.3205

Appendix 6. Descriptive Summary of the NKI70 Data

Total Gene Signatures = 70

Rows	Factors	n	Mean ± STD	Median	0.25	0.75	Range
36	PECI	144	-0.0071 ± 0.16976	0.0059	-0.0984	0.0864	-0.5563, 0.3561
37	MTDH	144	-0.0367 ± 0.16377	-0.0440	-0.1482	0.0588	-0.4336, 0.5128
38	Contig40831.RC	144	-0.0198 ± 0.20184	-0.0542	-0.1431	0.081	-0.5692, 0.4946
39	TGFB3	144	-0.0411 ± 0.24558	0.0015	-0.1917	0.1235	-0.6646, 0.4281
40	MELK	144	-0.0291 ± 0.22062	-0.0061	-0.1749	0.1357	-0.7679, 0.6030
41	COL4A2	144	-0.0007 ± 0.15617	0.0142	-0.1008	0.0897	-0.4826, 0.5083
42	DTL	144	-0.0223 ± 0.19384	0.0057	-0.1404	0.107	-0.5978, 0.4822
43	STK32B	144	0.0222 ± 0.13157	0.0082	-0.0466	0.0974	-0.3794, 0.5401
44	DCK	144	-0.0127 ± 0.15350	-0.0133	-0.1135	0.0714	-0.5321, 0.4563
45	FBXO31	144	-0.0258 ± 0.15446	-0.0254	-0.1291	0.0674	-0.5152, 0.4372
46	GPR126	144	-0.1129 ± 0.30481	-0.1240	-0.3329	0.0711	-0.8704, 0.7527
47	SLC2A3	144	-0.0275 ± 0.16628	-0.0475	-0.141	0.0834	-0.4215, 0.5556
48	PECI.1	144	0.0134 ± 0.16070	0.029	-0.0876	0.1256	-0.4404, 0.3755
49	ORC6L	144	-0.0656 ± 0.25671	-0.0504	-0.2841	0.1493	-0.5915, 0.5519
50	RFC4	144	-0.0267 ± 0.16663	-0.0485	-0.1319	0.0607	-0.6943, 0.5336
51	CDCA7	144	-0.3140 ± 0.38277	-0.3469	-0.5934	-0.1256	-1.0530, 0.7433
52	LOC643008	144	-0.0864 ± 0.31454	-0.1470	-0.3107	0.0977	-0.5852, 1.1000
53	MS4A7	144	-0.0378 ± 0.27792	-0.0546	-0.2388	0.112	-0.5971, 0.6692
54	MCM6	144	-0.0176 ± 0.17726	-0.0209	-0.1296	0.089	-0.6163, 0.4889
55	AP2B1	144	0.0151 ± 0.16976	0.039	-0.0907	0.1184	-0.5534, 0.5454
56	C9orf30	144	-0.0124 ± 0.12703	0.0034	-0.0995	0.0646	-0.4014, 0.3694
57	IGFBP5	144	-0.1427 ± 0.33692	-0.2206	-0.3652	0.0349	-0.6739, 1.1030
58	HRASLS	144	-0.0602 ± 0.23481	-0.0764	-0.173	0.0462	-1.1200, 0.7052
59	PITRM1	144	-0.0133 ± 0.15889	-0.0372	-0.1143	0.0835	-0.3872, 0.5575
60	IGFBP5.1	144	-0.0842 ± 0.29347	-0.1496	-0.2755	0.0337	-0.5351, 1.1440
61	NMU	144	-0.0512 ± 0.31639	-0.0896	-0.175	0.0495	-2.0000, 2.0000
62	PALM2.AKAP2	144	-0.0450 ± 0.18854	-0.0651	-0.162	0.0615	-0.6590, 0.7457
63	LGP2	144	0.0067 ± 0.18051	-0.0172	-0.1125	0.1293	-0.4010, 0.4013
64	PRC1	144	-0.0650 ± 0.25390	-0.0322	-0.2381	0.1156	-0.8477, 0.6077
65	Contig20217.RC	144	-0.0476 ± 0.20700	-0.0856	-0.1799	0.0504	-0.4088, 0.7331
66	CENPA	144	-0.1320 ± 0.31172	-0.1234	-0.3689	0.0844	-0.9092, 0.5376
67	EGLN1	144	-0.0224 ± 0.16088	-0.0144	-0.1298	0.0985	-0.5036, 0.4062
68	NM.004702	144	-0.0844 ± 0.27067	-0.0809	-0.2875	0.0931	-0.6677, 0.7101
69	ESM1	144	-0.0402 ± 0.26350	-0.0402	-0.1757	0.0895	-0.8972, 0.8474
70	C20orf46	144	-0.0857 ± 0.25265	-0.1333	-0.2561	0.0197	-0.4506, 0.9915

APPENDIX 7. Component Variance Matrix Obtained from Principal Component Analysis of the NKI70 Data

Total Components = 75

Components #	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard Deviation	4.21	2.37	2.04	1.8	1.75	1.7	1.43	1.42	1.35	1.3
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.24	0.31	0.37	0.41	0.45	0.49	0.52	0.54	0.57	0.59
Components #	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	Comp.19	Comp.20
Standard Deviation	1.28	1.24	1.22	1.18	1.15	1.11	1.09	1.06	1.01	0.99
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.61	0.63	0.65	0.67	0.69	0.71	0.72	0.74	0.75	0.76
Components #	Comp.21	Comp.22	Comp.23	Comp.24	Comp.25	Comp.26	Comp.27	Comp.28	Comp.29	Comp.30
Standard Deviation	0.97	0.95	0.93	0.9	0.89	0.88	0.86	0.84	0.82	0.81
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.78	0.79	0.8	0.81	0.82	0.83	0.84	0.85	0.86	0.87
Components #	Comp.31	Comp.32	Comp.33	Comp.34	Comp.35	Comp.36	Comp.37	Comp.38	Comp.39	Comp.40
Standard Deviation	0.78	0.76	0.74	0.71	0.71	0.68	0.67	0.65	0.63	0.6
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.88	0.88	0.89	0.9	0.9	0.91	0.92	0.92	0.93	0.93
Components #	Comp.41	Comp.42	Comp.43	Comp.44	Comp.45	Comp.46	Comp.47	Comp.48	Comp.49	Comp.50
Standard Deviation	0.59	0.58	0.56	0.54	0.53	0.51	0.5	0.49	0.47	0.46
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.94	0.94	0.95	0.95	0.95	0.96	0.96	0.96	0.97	0.97
Components #	Comp.51	Comp.52	Comp.53	Comp.54	Comp.55	Comp.56	Comp.57	Comp.58	Comp.59	Comp.60
Standard Deviation	0.45	0.43	0.41	0.4	0.39	0.38	0.36	0.36	0.35	0.34
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99
Components #	Comp.61	Comp.62	Comp.63	Comp.64	Comp.65	Comp.66	Comp.67	Comp.68	Comp.69	Comp.70
Standard Deviation	0.31	0.3	0.29	0.28	0.27	0.25	0.25	0.22	0.21	0.2
Proportion of Variance	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
Cumulative Proportion	0.99	0.99	0.99	0.99	0.99	1	1	1	1	1
Components #	Comp.71	Comp.72	Comp.73	Comp.74	Comp.75					
Standard Deviation	0.19	0.18	0.17	0.16	0.09					
Proportion of Variance	0.24	0.24	0.24	0.24	0.24					
Cumulative Proportion	1	1	1	1	1					

APPENDIX 8. AIC vs. Number of Components for Principal Component Cox Regression Model – NKI70 Data

Total Components = 70

Components	1	2	3	4	5	6	7	8	9	10
AIC	426.16	427.53	421.66	418.29	419.8	420.32	420.08	421.76	418.19	417.23
Components	11	12	13	14	15	16	17	18	19	20
AIC	411.99	413.66	414.92	416.53	414.38	404.36	406.3	406.66	404.66	401.5
Components	21	22	23	24	25	26	27	28	29	30
AIC	399.32	390.24	391.98	384.71	384.75	382.32	381.46	382.13	381.78	383.1
Components	31	32	33	34	35	36	37	38	39	40
AIC	384.13	385.94	387.93	388.99	389.48	388.56	389.76	391.44	392.88	394.88
Components	41	42	43	44	45	46	47	48	49	50
AIC	396.33	397.61	396.33	398.32	382.92	381.42	382.09	382.18	381.88	383.52
Components	51	52	53	54	55	56	57	58	59	60
AIC	385.45	381.96	383.95	382.64	370.33	372.09	363.49	339.74	341.08	343.02
Components	61	62	63	64	65	66	67	68	69	70
AIC	344.26	345.73	347.61	349.61	349.15	351.14	347.24	348.04	344.72	346.7

APPENDIX 9. Preselected Interactions from Deviance Test – NKI70 Data

Total Interactions = 508

#	V1	V2	Interactions	P-val[1]
1	Age	GNAZ	I(Age^1):GNAZ	0.0472
2	Age	GNAZ	I(Age^1)+I(Age^2):GNAZ	0.0472
3	Age	GNAZ	I(Age^1)+I(Age^2)+I(Age^3):GNAZ	0.0472
4	Age	Contig40831.RC	I(Age^1):Contig40831.RC	0.0289
5	Age	Contig40831.RC	I(Age^1)+I(Age^2):Contig40831.RC	0.0289
6	Age	Contig40831.RC	I(Age^1)+I(Age^2)+I(Age^3):Contig40831.RC	0.0289
7	Contig63649.RC	RUNDC1	I(Contig63649.RC^1):RUNDC1	0.007
8	Contig63649.RC	RUNDC1	I(Contig63649.RC^1)+I(Contig63649.RC^2):RUNDC1	0.007
9	Contig63649.RC	WISP1	I(Contig63649.RC^1):I(WISP1^1)	0.0333
10	Contig63649.RC	WISP1	I(Contig63649.RC^1)+I(Contig63649.RC^2):I(WISP1^1)+I(WISP1^2)	0.0333
11	Contig63649.RC	WISP1	I(Contig63649.RC^1):I(WISP1^1)+I(WISP1^2)+I(WISP1^3)	0.0333
12	Contig63649.RC	CDC42BPA	I(Contig63649.RC^1):I(CDC42BPA^1)	0.0468
13	Contig63649.RC	CDC42BPA	I(Contig63649.RC^1)+I(Contig63649.RC^2):I(CDC42BPA^1)+I(CDC42BPA^2)	0.0468
14	DIAPH3	QSCN6L1	DIAPH3:I(QSCN6L1^1)	0.0241
15	DIAPH3	QSCN6L1	DIAPH3:I(QSCN6L1^1)+I(QSCN6L1^2)	0.0241
16	DIAPH3	GMPS	DIAPH3:GMPS	0.0063
17	DIAPH3	GSTM3	DIAPH3:GSTM3	0.0101
18	DIAPH3	GPR180	DIAPH3:GPR180	0.0234
19	DIAPH3	UCHL5	DIAPH3:UCHL5	0.0467
20	DIAPH3	Contig40831.RC	DIAPH3:Contig40831.RC	0.0077
21	DIAPH3	MELK	DIAPH3:I(MELK^1)	0.0414
22	DIAPH3	MELK	DIAPH3:I(MELK^1)+I(MELK^2)	0.0414
23	DIAPH3	DTL	DIAPH3:I(DTL^1)	0.0212
24	DIAPH3	DTL	DIAPH3:I(DTL^1)+I(DTL^2)	0.0212
25	DIAPH3	ORC6L	DIAPH3:I(ORC6L^1)	0.0251
26	DIAPH3	ORC6L	DIAPH3:I(ORC6L^1)+I(ORC6L^2)	0.0251
27	DIAPH3	CDCA7	DIAPH3:CDCA7	0.0064
28	DIAPH3	MCM6	DIAPH3:I(MCM6^1)	0.0032
29	DIAPH3	MCM6	DIAPH3:I(MCM6^1)+I(MCM6^2)	0.0032
30	DIAPH3	HRASLS	DIAPH3:HRASLS	0.0127
31	DIAPH3	PITRM1	DIAPH3:I(PITRM1^1)	0.0112
32	DIAPH3	PITRM1	DIAPH3:I(PITRM1^1)+I(PITRM1^2)	0.0112
33	DIAPH3	PITRM1	DIAPH3:I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0112
34	DIAPH3	CENPA	DIAPH3:CENPA	0.0102
35	NUSAP1	DIAPH3.2	NUSAP1:I(DIAPH3.2^1)	0.0466

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
36	NUSAP1	DIAPH3.2	NUSAP1:I(DIAPH3.2^1)+I(DIAPH3.2^2)	0.0466
37	NUSAP1	RAB6B	NUSAP1:RAB6B	0.0436
38	NUSAP1	TGFB3	NUSAP1:I(TGFB3^1)	0.0322
39	NUSAP1	TGFB3	NUSAP1:I(TGFB3^1)+I(TGFB3^2)	0.0322
40	NUSAP1	CDCA7	NUSAP1:CDCA7	0.0258
41	NUSAP1	MCM6	NUSAP1:I(MCM6^1)	0.011
42	NUSAP1	MCM6	NUSAP1:I(MCM6^1)+I(MCM6^2)	0.011
43	NUSAP1	HRASLS	NUSAP1:HRASLS	0.0058
44	NUSAP1	CENPA	NUSAP1:CENPA	0.0384
45	AA555029.RC	FLT1	AA555029.RC:FLT1	0.0423
46	AA555029.RC	RTN4RL1	AA555029.RC:RTN4RL1	0.022
47	ALDH4A1	RTN4RL1	ALDH4A1:RTN4RL1	0.0301
48	ALDH4A1	DTL	ALDH4A1:I(DTL^1)	0.0326
49	ALDH4A1	DTL	ALDH4A1:I(DTL^1)+I(DTL^2)	0.0326
50	ALDH4A1	AP2B1	ALDH4A1:AP2B1	0.0119
51	QSCN6L1	DIAPH3.1	I(QSCN6L1^1):I(DIAPH3.1^1)	0.0066
52	QSCN6L1	DIAPH3.1	I(QSCN6L1^1)+I(QSCN6L1^2):I(DIAPH3.1^1)+I(DIAPH3.1^2)	0.0066
53	QSCN6L1	DIAPH3.2	I(QSCN6L1^1):I(DIAPH3.2^1)	0.0034
54	QSCN6L1	DIAPH3.2	I(QSCN6L1^1)+I(QSCN6L1^2):I(DIAPH3.2^1)+I(DIAPH3.2^2)	0.0034
55	QSCN6L1	C16orf61	I(QSCN6L1^1):C16orf61	0.0005
56	QSCN6L1	C16orf61	I(QSCN6L1^1)+I(QSCN6L1^2):C16orf61	0.0005
57	QSCN6L1	ECT2	I(QSCN6L1^1):ECT2	0.0031
58	QSCN6L1	ECT2	I(QSCN6L1^1)+I(QSCN6L1^2):ECT2	0.0031
59	QSCN6L1	KNTC2	I(QSCN6L1^1):I(KNTC2^1)	0.0002
60	QSCN6L1	KNTC2	I(QSCN6L1^1)+I(QSCN6L1^2):I(KNTC2^1)+I(KNTC2^2)	0.0002
61	QSCN6L1	SERF1A	I(QSCN6L1^1):SERF1A	0.0058
62	QSCN6L1	SERF1A	I(QSCN6L1^1)+I(QSCN6L1^2):SERF1A	0.0058
63	QSCN6L1	MTDH	I(QSCN6L1^1):MTDH	0.0048
64	QSCN6L1	MTDH	I(QSCN6L1^1)+I(QSCN6L1^2):MTDH	0.0048
65	QSCN6L1	Contig40831.RC	I(QSCN6L1^1):Contig40831.RC	<.0001
66	QSCN6L1	Contig40831.RC	I(QSCN6L1^1)+I(QSCN6L1^2):Contig40831.RC	<.0001
67	QSCN6L1	TGFB3	I(QSCN6L1^1):I(TGFB3^1)	0.0176
68	QSCN6L1	TGFB3	I(QSCN6L1^1)+I(QSCN6L1^2):I(TGFB3^1)+I(TGFB3^2)	0.0176
69	QSCN6L1	MELK	I(QSCN6L1^1):I(MELK^1)	<.0001
70	QSCN6L1	MELK	I(QSCN6L1^1)+I(QSCN6L1^2):I(MELK^1)+I(MELK^2)	<.0001

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
71	QSCN6L1	DTL	I(QSCN6L1^1):I(DTL^1)	0.0384
72	QSCN6L1	DTL	I(QSCN6L1^1)+I(QSCN6L1^2):I(DTL^1)+I(DTL^2)	0.0384
73	QSCN6L1	FBXO31	I(QSCN6L1^1):FBXO31	0.039
74	QSCN6L1	FBXO31	I(QSCN6L1^1)+I(QSCN6L1^2):FBXO31	0.039
75	QSCN6L1	ORC6L	I(QSCN6L1^1):I(ORC6L^1)	0.0024
76	QSCN6L1	ORC6L	I(QSCN6L1^1)+I(QSCN6L1^2):I(ORC6L^1)+I(ORC6L^2)	0.0024
77	QSCN6L1	RFC4	I(QSCN6L1^1):RFC4	0.0059
78	QSCN6L1	RFC4	I(QSCN6L1^1)+I(QSCN6L1^2):RFC4	0.0059
79	QSCN6L1	CDCA7	I(QSCN6L1^1):CDCA7	0.0005
80	QSCN6L1	CDCA7	I(QSCN6L1^1)+I(QSCN6L1^2):CDCA7	0.0005
81	QSCN6L1	MCM6	I(QSCN6L1^1):I(MCM6^1)	0.0028
82	QSCN6L1	MCM6	I(QSCN6L1^1)+I(QSCN6L1^2):I(MCM6^1)+I(MCM6^2)	0.0028
83	QSCN6L1	Contig20217.RC	I(QSCN6L1^1):Contig20217.RC	0.0035
84	QSCN6L1	Contig20217.RC	I(QSCN6L1^1)+I(QSCN6L1^2):Contig20217.RC	0.0035
85	QSCN6L1	CENPA	I(QSCN6L1^1):CENPA	0.0317
86	QSCN6L1	CENPA	I(QSCN6L1^1)+I(QSCN6L1^2):CENPA	0.0317
87	QSCN6L1	NM.004702	I(QSCN6L1^1):NM.004702	0.0001
88	QSCN6L1	NM.004702	I(QSCN6L1^1)+I(QSCN6L1^2):NM.004702	0.0001
89	FGF18	SCUBE2	I(FGF18^1):SCUBE2	0.0307
90	FGF18	SCUBE2	I(FGF18^1)+I(FGF18^2):SCUBE2	0.0307
91	FGF18	SCUBE2	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):SCUBE2	0.0307
92	FGF18	WISP1	I(FGF18^1):I(WISP1^1)	0.0192
93	FGF18	WISP1	I(FGF18^1)+I(FGF18^2):I(WISP1^1)+I(WISP1^2)	0.0192
94	FGF18	WISP1	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):I(WISP1^1)+I(WISP1^2)+I(WISP1^3)	0.0192
95	FGF18	TGFB3	I(FGF18^1):I(TGFB3^1)	0.0034
96	FGF18	TGFB3	I(FGF18^1)+I(FGF18^2):I(TGFB3^1)+I(TGFB3^2)	0.0034
97	FGF18	TGFB3	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):I(TGFB3^1)	0.0034
98	FGF18	STK32B	I(FGF18^1):STK32B	0.0324
99	FGF18	STK32B	I(FGF18^1)+I(FGF18^2):STK32B	0.0324
100	FGF18	STK32B	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):STK32B	0.0324
101	FGF18	DCK	I(FGF18^1):DCK	0.0337
102	FGF18	DCK	I(FGF18^1)+I(FGF18^2):DCK	0.0337
103	FGF18	DCK	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):DCK	0.0337
104	FGF18	CDCA7	I(FGF18^1):CDCA7	0.0203
105	FGF18	CDCA7	I(FGF18^1)+I(FGF18^2):CDCA7	0.0203

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
106	FGF18	CDCA7	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):CDCA7	0.0203
107	FGF18	MCM6	I(FGF18^1):I(MCM6^1)	0.0344
108	FGF18	MCM6	I(FGF18^1)+I(FGF18^2):I(MCM6^1)+I(MCM6^2)	0.0344
109	FGF18	MCM6	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):I(MCM6^1)	0.0344
110	FGF18	PITRM1	I(FGF18^1):I(PITRM1^1)	0.0014
111	FGF18	PITRM1	I(FGF18^1)+I(FGF18^2):I(PITRM1^1)+I(PITRM1^2)	0.0014
112	FGF18	PITRM1	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0014
113	FGF18	NMU	I(FGF18^1):NMU	0.0496
114	FGF18	NMU	I(FGF18^1)+I(FGF18^2):NMU	0.0496
115	FGF18	NMU	I(FGF18^1)+I(FGF18^2)+I(FGF18^3):NMU	0.0496
116	DIAPH3.1	DIAPH3.2	I(DIAPH3.1^1):I(DIAPH3.2^1)	0.0218
117	DIAPH3.1	DIAPH3.2	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(DIAPH3.2^1)+I(DIAPH3.2^2)	0.0218
118	DIAPH3.1	RUNDC1	I(DIAPH3.1^1):RUNDC1	0.0305
119	DIAPH3.1	RUNDC1	I(DIAPH3.1^1)+I(DIAPH3.1^2):RUNDC1	0.0305
120	DIAPH3.1	ECT2	I(DIAPH3.1^1):ECT2	0.0371
121	DIAPH3.1	ECT2	I(DIAPH3.1^1)+I(DIAPH3.1^2):ECT2	0.0371
122	DIAPH3.1	GMPS	I(DIAPH3.1^1):GMPS	0.0056
123	DIAPH3.1	GMPS	I(DIAPH3.1^1)+I(DIAPH3.1^2):GMPS	0.0056
124	DIAPH3.1	KNTC2	I(DIAPH3.1^1):I(KNTC2^1)	0.0111
125	DIAPH3.1	KNTC2	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(KNTC2^1)+I(KNTC2^2)	0.0111
126	DIAPH3.1	GSTM3	I(DIAPH3.1^1):GSTM3	0.0133
127	DIAPH3.1	GSTM3	I(DIAPH3.1^1)+I(DIAPH3.1^2):GSTM3	0.0133
128	DIAPH3.1	UCHL5	I(DIAPH3.1^1):UCHL5	0.0339
129	DIAPH3.1	UCHL5	I(DIAPH3.1^1)+I(DIAPH3.1^2):UCHL5	0.0339
130	DIAPH3.1	Contig40831.RC	I(DIAPH3.1^1):Contig40831.RC	0.0007
131	DIAPH3.1	Contig40831.RC	I(DIAPH3.1^1)+I(DIAPH3.1^2):Contig40831.RC	0.0007
132	DIAPH3.1	TGFB3	I(DIAPH3.1^1):I(TGFB3^1)	0.0032
133	DIAPH3.1	TGFB3	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(TGFB3^1)+I(TGFB3^2)	0.0032
134	DIAPH3.1	MELK	I(DIAPH3.1^1):I(MELK^1)	0.036
135	DIAPH3.1	MELK	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(MELK^1)+I(MELK^2)	0.036
136	DIAPH3.1	DTL	I(DIAPH3.1^1):I(DTL^1)	0.0277
137	DIAPH3.1	DTL	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(DTL^1)+I(DTL^2)	0.0277
138	DIAPH3.1	GPR126	I(DIAPH3.1^1):GPR126	0.0322
139	DIAPH3.1	GPR126	I(DIAPH3.1^1)+I(DIAPH3.1^2):GPR126	0.0322
140	DIAPH3.1	ORC6L	I(DIAPH3.1^1):I(ORC6L^1)	0.0066

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
141	DIAPH3.1	ORC6L	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(ORC6L^1)+I(ORC6L^2)	0.0066
142	DIAPH3.1	RFC4	I(DIAPH3.1^1):RFC4	0.0271
143	DIAPH3.1	RFC4	I(DIAPH3.1^1)+I(DIAPH3.1^2):RFC4	0.0271
144	DIAPH3.1	CDCA7	I(DIAPH3.1^1):CDCA7	0.0013
145	DIAPH3.1	CDCA7	I(DIAPH3.1^1)+I(DIAPH3.1^2):CDCA7	0.0013
146	DIAPH3.1	MCM6	I(DIAPH3.1^1):I(MCM6^1)	0.0002
147	DIAPH3.1	MCM6	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(MCM6^1)+I(MCM6^2)	0.0002
148	DIAPH3.1	PITRM1	I(DIAPH3.1^1):I(PITRM1^1)	0.0445
149	DIAPH3.1	PITRM1	I(DIAPH3.1^1)+I(DIAPH3.1^2):I(PITRM1^1)+I(PITRM1^2)	0.0445
150	DIAPH3.1	PITRM1	I(DIAPH3.1^1):I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0445
151	DIAPH3.1	NMU	I(DIAPH3.1^1):NMU	0.0248
152	DIAPH3.1	NMU	I(DIAPH3.1^1)+I(DIAPH3.1^2):NMU	0.0248
153	DIAPH3.1	Contig20217.RC	I(DIAPH3.1^1):Contig20217.RC	0.0061
154	DIAPH3.1	Contig20217.RC	I(DIAPH3.1^1)+I(DIAPH3.1^2):Contig20217.RC	0.0061
155	DIAPH3.1	CENPA	I(DIAPH3.1^1):CENPA	0.0018
156	DIAPH3.1	CENPA	I(DIAPH3.1^1)+I(DIAPH3.1^2):CENPA	0.0018
157	Contig32125.RC	MTDH	I(Contig32125.RC^1):MTDH	0.0215
158	Contig32125.RC	MTDH	I(Contig32125.RC^1)+I(Contig32125.RC^2):MTDH	0.0215
159	Contig32125.RC	MTDH	I(Contig32125.RC^1)+I(Contig32125.RC^2)+I(Contig32125.RC^3):MTDH	0.0215
160	Contig32125.RC	CDCA7	I(Contig32125.RC^1):CDCA7	0.0054
161	Contig32125.RC	CDCA7	I(Contig32125.RC^1)+I(Contig32125.RC^2):CDCA7	0.0054
162	Contig32125.RC	CDCA7	I(Contig32125.RC^1)+I(Contig32125.RC^2)+I(Contig32125.RC^3):CDCA7	0.0054
163	Contig32125.RC	MCM6	I(Contig32125.RC^1):I(MCM6^1)	0.0062
164	Contig32125.RC	MCM6	I(Contig32125.RC^1)+I(Contig32125.RC^2):I(MCM6^1)+I(MCM6^2)	0.0062
165	Contig32125.RC	MCM6	I(Contig32125.RC^1)+I(Contig32125.RC^2)+I(Contig32125.RC^3):I(MCM6^1)	0.0062
166	BBC3	KNTC2	BBC3:I(KNTC2^1)	0.0453
167	BBC3	KNTC2	BBC3:I(KNTC2^1)+I(KNTC2^2)	0.0453
168	BBC3	AYTL2	BBC3:AYTL2	0.0407
169	BBC3	Contig40831.RC	BBC3:Contig40831.RC	0.0288
170	BBC3	MS4A7	BBC3:MS4A7	0.0292
171	BBC3	IGFBP5	BBC3:IGFBP5	0.0473
172	BBC3	IGFBP5.1	BBC3:IGFBP5.1	0.0174
173	BBC3	PALM2.AKAP2	BBC3:PALM2.AKAP2	0.036
174	BBC3	Contig20217.RC	BBC3:Contig20217.RC	0.0396
175	DIAPH3.2	C16orf61	I(DIAPH3.2^1):C16orf61	0.0254

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
176	DIAPH3.2	C16orf61	I(DIAPH3.2^1)+I(DIAPH3.2^2):C16orf61	0.0254
177	DIAPH3.2	ECT2	I(DIAPH3.2^1):ECT2	0.0107
178	DIAPH3.2	ECT2	I(DIAPH3.2^1)+I(DIAPH3.2^2):ECT2	0.0107
179	DIAPH3.2	GMPS	I(DIAPH3.2^1):GMPS	0.0115
180	DIAPH3.2	GMPS	I(DIAPH3.2^1)+I(DIAPH3.2^2):GMPS	0.0115
181	DIAPH3.2	KNTC2	I(DIAPH3.2^1):I(KNTC2^1)	0.0238
182	DIAPH3.2	KNTC2	I(DIAPH3.2^1)+I(DIAPH3.2^2):I(KNTC2^1)+I(KNTC2^2)	0.0238
183	DIAPH3.2	UCHL5	I(DIAPH3.2^1):UCHL5	0.0311
184	DIAPH3.2	UCHL5	I(DIAPH3.2^1)+I(DIAPH3.2^2):UCHL5	0.0311
185	DIAPH3.2	Contig40831.RC	I(DIAPH3.2^1):Contig40831.RC	0.0034
186	DIAPH3.2	Contig40831.RC	I(DIAPH3.2^1)+I(DIAPH3.2^2):Contig40831.RC	0.0034
187	DIAPH3.2	TGFB3	I(DIAPH3.2^1):I(TGFB3^1)	0.0118
188	DIAPH3.2	TGFB3	I(DIAPH3.2^1)+I(DIAPH3.2^2):I(TGFB3^1)+I(TGFB3^2)	0.0118
189	DIAPH3.2	MELK	I(DIAPH3.2^1):I(MELK^1)	0.0148
190	DIAPH3.2	MELK	I(DIAPH3.2^1)+I(DIAPH3.2^2):I(MELK^1)+I(MELK^2)	0.0148
191	DIAPH3.2	DTL	I(DIAPH3.2^1):I(DTL^1)	0.0106
192	DIAPH3.2	DTL	I(DIAPH3.2^1)+I(DIAPH3.2^2):I(DTL^1)+I(DTL^2)	0.0106
193	DIAPH3.2	ORC6L	I(DIAPH3.2^1):I(ORC6L^1)	0.0033
194	DIAPH3.2	ORC6L	I(DIAPH3.2^1)+I(DIAPH3.2^2):I(ORC6L^1)+I(ORC6L^2)	0.0033
195	DIAPH3.2	CDCA7	I(DIAPH3.2^1):CDCA7	0.0026
196	DIAPH3.2	CDCA7	I(DIAPH3.2^1)+I(DIAPH3.2^2):CDCA7	0.0026
197	DIAPH3.2	MCM6	I(DIAPH3.2^1):I(MCM6^1)	0.0001
198	DIAPH3.2	MCM6	I(DIAPH3.2^1)+I(DIAPH3.2^2):I(MCM6^1)+I(MCM6^2)	0.0001
199	DIAPH3.2	LGP2	I(DIAPH3.2^1):LGP2	0.0406
200	DIAPH3.2	LGP2	I(DIAPH3.2^1)+I(DIAPH3.2^2):LGP2	0.0406
201	DIAPH3.2	Contig20217.RC	I(DIAPH3.2^1):Contig20217.RC	0.0315
202	DIAPH3.2	Contig20217.RC	I(DIAPH3.2^1)+I(DIAPH3.2^2):Contig20217.RC	0.0315
203	DIAPH3.2	CENPA	I(DIAPH3.2^1):CENPA	0.0005
204	DIAPH3.2	CENPA	I(DIAPH3.2^1)+I(DIAPH3.2^2):CENPA	0.0005
205	DIAPH3.2	NM.004702	I(DIAPH3.2^1):NM.004702	0.0293
206	DIAPH3.2	NM.004702	I(DIAPH3.2^1)+I(DIAPH3.2^2):NM.004702	0.0293
207	RP5.860F19.3	Contig40831.RC	I(RP5.860F19.3^1):Contig40831.RC	0.0076
208	RP5.860F19.3	Contig40831.RC	I(RP5.860F19.3^1)+I(RP5.860F19.3^2):Contig40831.RC	0.0076
209	RP5.860F19.3	MELK	I(RP5.860F19.3^1):I(MELK^1)	0.0413
210	RP5.860F19.3	MELK	I(RP5.860F19.3^1)+I(RP5.860F19.3^2):I(MELK^1)+I(MELK^2)	0.0413

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
211	RP5.860F19.3	RFC4	I(RP5.860F19.3^1):RFC4	0.041
212	RP5.860F19.3	RFC4	I(RP5.860F19.3^1)+I(RP5.860F19.3^2):RFC4	0.041
213	RP5.860F19.3	EGLN1	I(RP5.860F19.3^1):EGLN1	0.0341
214	RP5.860F19.3	EGLN1	I(RP5.860F19.3^1)+I(RP5.860F19.3^2):EGLN1	0.0341
215	RP5.860F19.3	NM.004702	I(RP5.860F19.3^1):NM.004702	0.0196
216	RP5.860F19.3	NM.004702	I(RP5.860F19.3^1)+I(RP5.860F19.3^2):NM.004702	0.0196
217	C16orf61	GMPS	C16orf61:GMPS	0.0243
218	C16orf61	TGFB3	C16orf61:I(TGFB3^1)	0.0132
219	C16orf61	TGFB3	C16orf61:I(TGFB3^1)+I(TGFB3^2)	0.0132
220	C16orf61	MELK	C16orf61:I(MELK^1)	0.0246
221	C16orf61	MELK	C16orf61:I(MELK^1)+I(MELK^2)	0.0246
222	C16orf61	DTL	C16orf61:I(DTL^1)	0.0378
223	C16orf61	DTL	C16orf61:I(DTL^1)+I(DTL^2)	0.0378
224	C16orf61	CDCA7	C16orf61:CDCA7	0.0103
225	C16orf61	MCM6	C16orf61:I(MCM6^1)	0.0038
226	C16orf61	MCM6	C16orf61:I(MCM6^1)+I(MCM6^2)	0.0038
227	C16orf61	HRASLS	C16orf61:HRASLS	0.029
228	SCUBE2	Contig35251.RC	SCUBE2:Contig35251.RC	0.0145
229	SCUBE2	AYTL2	SCUBE2:AYTL2	0.0494
230	SCUBE2	RAB6B	SCUBE2:RAB6B	0.0484
231	SCUBE2	PECI	SCUBE2:PECI	0.0489
232	SCUBE2	TGFB3	SCUBE2:I(TGFB3^1)	0.0362
233	SCUBE2	TGFB3	SCUBE2:I(TGFB3^1)+I(TGFB3^2)	0.0362
234	SCUBE2	STK32B	SCUBE2:STK32B	0.0131
235	SCUBE2	PECI.1	SCUBE2:PECI.1	0.0349
236	SCUBE2	LOC643008	SCUBE2:I(LOC643008^1)	0.0271
237	SCUBE2	LOC643008	SCUBE2:I(LOC643008^1)+I(LOC643008^2)	0.0271
238	SCUBE2	LOC643008	SCUBE2:I(LOC643008^1)+I(LOC643008^2)+I(LOC643008^3)	0.0271
239	FLT1	MMP9	FLT1:MMP9	0.0079
240	FLT1	RUNDC1	FLT1:RUNDC1	0.0075
241	FLT1	Contig35251.RC	FLT1:Contig35251.RC	0.0232
242	FLT1	WISP1	FLT1:I(WISP1^1)	0.0071
243	FLT1	WISP1	FLT1:I(WISP1^1)+I(WISP1^2)	0.0071
244	FLT1	WISP1	FLT1:I(WISP1^1)+I(WISP1^2)+I(WISP1^3)	0.0071
245	FLT1	AP2B1	FLT1:AP2B1	0.0171

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
246	FLT1	EGLN1	FLT1:EGLN1	0.0056
247	GNAZ	CDCA7	GNAZ:CDCA7	0.0372
248	GNAZ	MS4A7	GNAZ:MS4A7	0.0361
249	GNAZ	LGP2	GNAZ:LGP2	0.0011
250	OXCT1	IGFBP5.1	OXCT1:IGFBP5.1	0.0437
251	OXCT1	NMU	OXCT1:NMU	0.0131
252	MMP9	CDC42BPA	MMP9:I(CDC42BPA^1)	0.0285
253	MMP9	CDC42BPA	MMP9:I(CDC42BPA^1)+I(CDC42BPA^2)	0.0285
254	MMP9	COL4A2	MMP9:COL4A2	0.0368
255	MMP9	IGFBP5	MMP9:IGFBP5	0.0309
256	MMP9	IGFBP5.1	MMP9:IGFBP5.1	0.0444
257	RUNDC1	Contig35251.RC	RUNDC1:Contig35251.RC	0.038
258	RUNDC1	ECT2	RUNDC1:ECT2	0.0439
259	RUNDC1	TGFB3	RUNDC1:I(TGFB3^1)	0.0359
260	RUNDC1	TGFB3	RUNDC1:I(TGFB3^1)+I(TGFB3^2)	0.0359
261	RUNDC1	LGP2	RUNDC1:LGP2	0.0131
262	Contig35251.RC	ECT2	Contig35251.RC:ECT2	0.0195
263	Contig35251.RC	KNTC2	Contig35251.RC:I(KNTC2^1)	0.0277
264	Contig35251.RC	KNTC2	Contig35251.RC:I(KNTC2^1)+I(KNTC2^2)	0.0277
265	Contig35251.RC	CDC42BPA	Contig35251.RC:I(CDC42BPA^1)	0.0205
266	Contig35251.RC	CDC42BPA	Contig35251.RC:I(CDC42BPA^1)+I(CDC42BPA^2)	0.0205
267	Contig35251.RC	RAB6B	Contig35251.RC:RAB6B	0.0364
268	Contig35251.RC	ZNF533	Contig35251.RC:ZNF533	0.046
269	Contig35251.RC	MTDH	Contig35251.RC:MTDH	0.0241
270	Contig35251.RC	Contig40831.RC	Contig35251.RC:Contig40831.RC	0.0041
271	Contig35251.RC	TGFB3	Contig35251.RC:I(TGFB3^1)	0.0068
272	Contig35251.RC	TGFB3	Contig35251.RC:I(TGFB3^1)+I(TGFB3^2)	0.0068
273	Contig35251.RC	ORC6L	Contig35251.RC:I(ORC6L^1)	0.0432
274	Contig35251.RC	ORC6L	Contig35251.RC:I(ORC6L^1)+I(ORC6L^2)	0.0432
275	Contig35251.RC	CENPA	Contig35251.RC:CENPA	0.0379
276	Contig35251.RC	NM.004702	Contig35251.RC:NM.004702	0.032
277	ECT2	KNTC2	ECT2:I(KNTC2^1)	0.0141
278	ECT2	KNTC2	ECT2:I(KNTC2^1)+I(KNTC2^2)	0.0141
279	ECT2	UCHL5	ECT2:UCHL5	0.0422
280	ECT2	Contig40831.RC	ECT2:Contig40831.RC	0.0405

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
281	ECT2	TGFB3	ECT2:I(TGFB3^1)	0.0003
282	ECT2	TGFB3	ECT2:I(TGFB3^1)+I(TGFB3^2)	0.0003
283	ECT2	MELK	ECT2:I(MELK^1)	0.0493
284	ECT2	MELK	ECT2:I(MELK^1)+I(MELK^2)	0.0493
285	ECT2	DTL	ECT2:I(DTL^1)	0.0242
286	ECT2	DTL	ECT2:I(DTL^1)+I(DTL^2)	0.0242
287	ECT2	DCK	ECT2:DCK	0.0272
288	ECT2	GPR126	ECT2:GPR126	0.0001
289	ECT2	CDCA7	ECT2:CDCA7	0.0001
290	ECT2	MCM6	ECT2:I(MCM6^1)	0.0001
291	ECT2	MCM6	ECT2:I(MCM6^1)+I(MCM6^2)	0.0001
292	ECT2	HRASLS	ECT2:HRASLS	0.0291
293	ECT2	CENPA	ECT2:CENPA	0.0286
294	GMPS	KNTC2	GMPS:I(KNTC2^1)	0.021
295	GMPS	KNTC2	GMPS:I(KNTC2^1)+I(KNTC2^2)	0.021
296	GMPS	DTL	GMPS:I(DTL^1)	0.0235
297	GMPS	DTL	GMPS:I(DTL^1)+I(DTL^2)	0.0235
298	GMPS	CDCA7	GMPS:CDCA7	0.0122
299	GMPS	MCM6	GMPS:I(MCM6^1)	0.023
300	GMPS	MCM6	GMPS:I(MCM6^1)+I(MCM6^2)	0.023
301	GMPS	PRC1	GMPS:PRC1	0.047
302	GMPS	Contig20217.RC	GMPS:Contig20217.RC	0.0443
303	GMPS	EGLN1	GMPS:EGLN1	0.0298
304	KNTC2	RAB6B	I(KNTC2^1):RAB6B	0.0048
305	KNTC2	RAB6B	I(KNTC2^1)+I(KNTC2^2):RAB6B	0.0048
306	KNTC2	Contig40831.RC	I(KNTC2^1):Contig40831.RC	0.0084
307	KNTC2	Contig40831.RC	I(KNTC2^1)+I(KNTC2^2):Contig40831.RC	0.0084
308	KNTC2	TGFB3	I(KNTC2^1):I(TGFB3^1)	0.0007
309	KNTC2	TGFB3	I(KNTC2^1)+I(KNTC2^2):I(TGFB3^1)+I(TGFB3^2)	0.0007
310	KNTC2	MELK	I(KNTC2^1):I(MELK^1)	0.0148
311	KNTC2	MELK	I(KNTC2^1)+I(KNTC2^2):I(MELK^1)+I(MELK^2)	0.0148
312	KNTC2	GPR126	I(KNTC2^1):GPR126	0.0407
313	KNTC2	GPR126	I(KNTC2^1)+I(KNTC2^2):GPR126	0.0407
314	KNTC2	PECI.1	I(KNTC2^1):PECI.1	0.0475
315	KNTC2	PECI.1	I(KNTC2^1)+I(KNTC2^2):PECI.1	0.0475

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
316	KNTC2	CDCA7	I(KNTC2^1):CDCA7	0.0034
317	KNTC2	CDCA7	I(KNTC2^1)+I(KNTC2^2):CDCA7	0.0034
318	KNTC2	LOC643008	I(KNTC2^1):I(LOC643008^1)	0.021
319	KNTC2	LOC643008	I(KNTC2^1)+I(KNTC2^2):I(LOC643008^1)+I(LOC643008^2)	0.021
320	KNTC2	LOC643008	I(KNTC2^1):I(LOC643008^1)+I(LOC643008^2)+I(LOC643008^3)	0.021
321	KNTC2	IGFBP5	I(KNTC2^1):IGFBP5	0.0097
322	KNTC2	IGFBP5	I(KNTC2^1)+I(KNTC2^2):IGFBP5	0.0097
323	KNTC2	IGFBP5.1	I(KNTC2^1):IGFBP5.1	0.0063
324	KNTC2	IGFBP5.1	I(KNTC2^1)+I(KNTC2^2):IGFBP5.1	0.0063
325	KNTC2	Contig20217.RC	I(KNTC2^1):Contig20217.RC	0.0054
326	KNTC2	Contig20217.RC	I(KNTC2^1)+I(KNTC2^2):Contig20217.RC	0.0054
327	KNTC2	NM.004702	I(KNTC2^1):NM.004702	0.0295
328	KNTC2	NM.004702	I(KNTC2^1)+I(KNTC2^2):NM.004702	0.0295
329	WISP1	UCHL5	I(WISP1^1):UCHL5	0.0107
330	WISP1	UCHL5	I(WISP1^1)+I(WISP1^2):UCHL5	0.0107
331	WISP1	UCHL5	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):UCHL5	0.0107
332	WISP1	DTL	I(WISP1^1):I(DTL^1)	0.0499
333	WISP1	DTL	I(WISP1^1)+I(WISP1^2):I(DTL^1)+I(DTL^2)	0.0499
334	WISP1	DTL	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):I(DTL^1)	0.0499
335	WISP1	STK32B	I(WISP1^1):STK32B	0.0061
336	WISP1	STK32B	I(WISP1^1)+I(WISP1^2):STK32B	0.0061
337	WISP1	STK32B	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):STK32B	0.0061
338	WISP1	ORC6L	I(WISP1^1):I(ORC6L^1)	0.0241
339	WISP1	ORC6L	I(WISP1^1)+I(WISP1^2):I(ORC6L^1)+I(ORC6L^2)	0.0241
340	WISP1	ORC6L	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):I(ORC6L^1)	0.0241
341	WISP1	CDCA7	I(WISP1^1):CDCA7	0.0306
342	WISP1	CDCA7	I(WISP1^1)+I(WISP1^2):CDCA7	0.0306
343	WISP1	CDCA7	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):CDCA7	0.0306
344	WISP1	MCM6	I(WISP1^1):I(MCM6^1)	0.0342
345	WISP1	MCM6	I(WISP1^1)+I(WISP1^2):I(MCM6^1)+I(MCM6^2)	0.0342
346	WISP1	MCM6	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):I(MCM6^1)	0.0342
347	WISP1	PITRM1	I(WISP1^1):I(PITRM1^1)	0.0142
348	WISP1	PITRM1	I(WISP1^1)+I(WISP1^2):I(PITRM1^1)+I(PITRM1^2)	0.0142
349	WISP1	PITRM1	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0142
350	WISP1	PRC1	I(WISP1^1):PRC1	0.0048

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

Total Interactions = 508

#	V1	V2	Interactions	P-val[1]
351	WISP1	PRC1	I(WISP1^1)+I(WISP1^2):PRC1	0.0048
352	WISP1	PRC1	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):PRC1	0.0048
353	WISP1	ESM1	I(WISP1^1):ESM1	0.0326
354	WISP1	ESM1	I(WISP1^1)+I(WISP1^2):ESM1	0.0326
355	WISP1	ESM1	I(WISP1^1)+I(WISP1^2)+I(WISP1^3):ESM1	0.0326
356	CDC42BPA	LOC643008	I(CDC42BPA^1):I(LOC643008^1)	0.003
357	CDC42BPA	LOC643008	I(CDC42BPA^1)+I(CDC42BPA^2):I(LOC643008^1)+I(LOC643008^2)	0.003
358	CDC42BPA	LOC643008	I(CDC42BPA^1):I(LOC643008^1)+I(LOC643008^2)+I(LOC643008^3)	0.003
359	SERF1A	CDCA7	SERF1A:CDCA7	0.0312
360	SERF1A	MCM6	SERF1A:I(MCM6^1)	0.0208
361	SERF1A	MCM6	SERF1A:I(MCM6^1)+I(MCM6^2)	0.0208
362	SERF1A	NMU	SERF1A:NMU	0.0417
363	AYTL2	RTN4RL1	AYTL2:RTN4RL1	0.0356
364	AYTL2	STK32B	AYTL2:STK32B	0.0449
365	AYTL2	MS4A7	AYTL2:MS4A7	0.0008
366	AYTL2	C9orf30	AYTL2:C9orf30	0.0481
367	AYTL2	NMU	AYTL2:NMU	0.0106
368	GSTM3	IGFBP5.1	GSTM3:IGFBP5.1	0.0402
369	GSTM3	LGP2	GSTM3:LGP2	0.0045
370	GPR180	MELK	GPR180:I(MELK^1)	0.031
371	GPR180	MELK	GPR180:I(MELK^1)+I(MELK^2)	0.031
372	RAB6B	UCHL5	RAB6B:UCHL5	0.0409
373	RAB6B	Contig40831.RC	RAB6B:Contig40831.RC	0.0395
374	RAB6B	ORC6L	RAB6B:I(ORC6L^1)	0.0209
375	RAB6B	ORC6L	RAB6B:I(ORC6L^1)+I(ORC6L^2)	0.0209
376	RAB6B	RFC4	RAB6B:RFC4	0.0129
377	RAB6B	CDCA7	RAB6B:CDCA7	0.008
378	RAB6B	C9orf30	RAB6B:C9orf30	0.02
379	RAB6B	PRC1	RAB6B:PRC1	0.0045
380	RAB6B	Contig20217.RC	RAB6B:Contig20217.RC	0.0032
381	RAB6B	EGLN1	RAB6B:EGLN1	0.0071
382	RTN4RL1	HRASLS	RTN4RL1:HRASLS	0.0389
383	RTN4RL1	PALM2.AKAP2	RTN4RL1:PALM2.AKAP2	0.0414
384	UCHL5	Contig40831.RC	UCHL5:Contig40831.RC	0.0425
385	UCHL5	TGFB3	UCHL5:I(TGFB3^1)	0.0097

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

Total Interactions = 508

#	V1	V2	Interactions	P-val[1]
386	UCHL5	TGFB3	UCHL5:I(TGFB3^1)+I(TGFB3^2)	0.0097
387	UCHL5	MS4A7	UCHL5:MS4A7	0.0106
388	PECI	PALM2.AKAP2	PECI:PALM2.AKAP2	0.0234
389	MTDH	CDCA7	MTDH:CDCA7	0.0106
390	MTDH	PITRM1	MTDH:I(PITRM1^1)	0.0186
391	MTDH	PITRM1	MTDH:I(PITRM1^1)+I(PITRM1^2)	0.0186
392	MTDH	PITRM1	MTDH:I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0186
393	MTDH	EGLN1	MTDH:EGLN1	0.032
394	Contig40831.RC	TGFB3	Contig40831.RC:I(TGFB3^1)	0.0007
395	Contig40831.RC	TGFB3	Contig40831.RC:I(TGFB3^1)+I(TGFB3^2)	0.0007
396	Contig40831.RC	MELK	Contig40831.RC:I(MELK^1)	0.0037
397	Contig40831.RC	MELK	Contig40831.RC:I(MELK^1)+I(MELK^2)	0.0037
398	Contig40831.RC	STK32B	Contig40831.RC:STK32B	0.0459
399	Contig40831.RC	ORC6L	Contig40831.RC:I(ORC6L^1)	0.0262
400	Contig40831.RC	ORC6L	Contig40831.RC:I(ORC6L^1)+I(ORC6L^2)	0.0262
401	Contig40831.RC	CDCA7	Contig40831.RC:CDCA7	0.0172
402	Contig40831.RC	MCM6	Contig40831.RC:I(MCM6^1)	0.0065
403	Contig40831.RC	MCM6	Contig40831.RC:I(MCM6^1)+I(MCM6^2)	0.0065
404	Contig40831.RC	PITRM1	Contig40831.RC:I(PITRM1^1)	0.0246
405	Contig40831.RC	PITRM1	Contig40831.RC:I(PITRM1^1)+I(PITRM1^2)	0.0246
406	Contig40831.RC	PITRM1	Contig40831.RC:I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0246
407	Contig40831.RC	CENPA	Contig40831.RC:CENPA	0.021
408	Contig40831.RC	NM.004702	Contig40831.RC:NM.004702	0.0426
409	TGFB3	MELK	I(TGFB3^1):I(MELK^1)	0.0212
410	TGFB3	MELK	I(TGFB3^1)+I(TGFB3^2):I(MELK^1)+I(MELK^2)	0.0212
411	TGFB3	DCK	I(TGFB3^1):DCK	0.0281
412	TGFB3	DCK	I(TGFB3^1)+I(TGFB3^2):DCK	0.0281
413	TGFB3	ORC6L	I(TGFB3^1):I(ORC6L^1)	0.0392
414	TGFB3	ORC6L	I(TGFB3^1)+I(TGFB3^2):I(ORC6L^1)+I(ORC6L^2)	0.0392
415	TGFB3	RFC4	I(TGFB3^1):RFC4	0.0407
416	TGFB3	RFC4	I(TGFB3^1)+I(TGFB3^2):RFC4	0.0407
417	TGFB3	CDCA7	I(TGFB3^1):CDCA7	0.0268
418	TGFB3	CDCA7	I(TGFB3^1)+I(TGFB3^2):CDCA7	0.0268
419	TGFB3	MCM6	I(TGFB3^1):I(MCM6^1)	0.0178
420	TGFB3	MCM6	I(TGFB3^1)+I(TGFB3^2):I(MCM6^1)+I(MCM6^2)	0.0178

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

Total Interactions = 508

#	V1	V2	Interactions	P-val[1]
421	TGFB3	AP2B1	I(TGFB3^1):AP2B1	0.0349
422	TGFB3	AP2B1	I(TGFB3^1)+I(TGFB3^2):AP2B1	0.0349
423	TGFB3	PITRM1	I(TGFB3^1):I(PITRM1^1)	0.0039
424	TGFB3	PITRM1	I(TGFB3^1)+I(TGFB3^2):I(PITRM1^1)+I(PITRM1^2)	0.0039
425	TGFB3	PITRM1	I(TGFB3^1):I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0039
426	TGFB3	IGFBP5.1	I(TGFB3^1):IGFBP5.1	0.0233
427	TGFB3	IGFBP5.1	I(TGFB3^1)+I(TGFB3^2):IGFBP5.1	0.0233
428	TGFB3	Contig20217.RC	I(TGFB3^1):Contig20217.RC	0.0067
429	TGFB3	Contig20217.RC	I(TGFB3^1)+I(TGFB3^2):Contig20217.RC	0.0067
430	TGFB3	NM.004702	I(TGFB3^1):NM.004702	0.0227
431	TGFB3	NM.004702	I(TGFB3^1)+I(TGFB3^2):NM.004702	0.0227
432	MELK	RFC4	I(MELK^1):RFC4	0.0479
433	MELK	RFC4	I(MELK^1)+I(MELK^2):RFC4	0.0479
434	MELK	CDCA7	I(MELK^1):CDCA7	0.0077
435	MELK	CDCA7	I(MELK^1)+I(MELK^2):CDCA7	0.0077
436	MELK	MCM6	I(MELK^1):I(MCM6^1)	0.0023
437	MELK	MCM6	I(MELK^1)+I(MELK^2):I(MCM6^1)+I(MCM6^2)	0.0023
438	MELK	PITRM1	I(MELK^1):I(PITRM1^1)	0.034
439	MELK	PITRM1	I(MELK^1)+I(MELK^2):I(PITRM1^1)+I(PITRM1^2)	0.034
440	MELK	PITRM1	I(MELK^1):I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.034
441	MELK	CENPA	I(MELK^1):CENPA	0.0028
442	MELK	CENPA	I(MELK^1)+I(MELK^2):CENPA	0.0028
443	COL4A2	HRASLS	COL4A2:HRASLS	0.0287
444	COL4A2	LGP2	COL4A2:LGP2	0.0131
445	DTL	MCM6	I(DTL^1):I(MCM6^1)	0.0085
446	DTL	MCM6	I(DTL^1)+I(DTL^2):I(MCM6^1)+I(MCM6^2)	0.0085
447	DTL	NMU	I(DTL^1):NMU	0.041
448	DTL	NMU	I(DTL^1)+I(DTL^2):NMU	0.041
449	STK32B	FBXO31	STK32B:FBXO31	0.0356
450	STK32B	PRC1	STK32B:PRC1	0.0114
451	DCK	GPR126	DCK:GPR126	0.0044
452	DCK	SLC2A3	DCK:SLC2A3	0.0222
453	DCK	NM.004702	DCK:NM.004702	0.0427
454	FBXO31	HRASLS	FBXO31:HRASLS	0.0018
455	FBXO31	PITRM1	FBXO31:I(PITRM1^1)	0.0322

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

#	V1	V2	Interactions	P-val[1]
456	FBXO31	PITRM1	FBXO31:I(PITRM1^1)+I(PITRM1^2)	0.0322
457	FBXO31	PITRM1	FBXO31:I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0322
458	SLC2A3	CDCA7	SLC2A3:CDCA7	0.0423
459	SLC2A3	LGP2	SLC2A3:LGP2	0.0292
460	PECI.1	RFC4	PECI.1:RFC4	0.0382
461	PECI.1	PALM2.AKAP2	PECI.1:PALM2.AKAP2	0.0423
462	ORC6L	LOC643008	I(ORC6L^1):I(LOC643008^1)	0.0162
463	ORC6L	LOC643008	I(ORC6L^1)+I(ORC6L^2):I(LOC643008^1)+I(LOC643008^2)	0.0162
464	ORC6L	LOC643008	I(ORC6L^1):I(LOC643008^1)+I(LOC643008^2)+I(LOC643008^3)	0.0162
465	ORC6L	Contig20217.RC	I(ORC6L^1):Contig20217.RC	0.0295
466	ORC6L	Contig20217.RC	I(ORC6L^1)+I(ORC6L^2):Contig20217.RC	0.0295
467	ORC6L	CENPA	I(ORC6L^1):CENPA	0.0235
468	ORC6L	CENPA	I(ORC6L^1)+I(ORC6L^2):CENPA	0.0235
469	RFC4	CDCA7	RFC4:CDCA7	0.0083
470	RFC4	MCM6	RFC4:I(MCM6^1)	0.0193
471	RFC4	MCM6	RFC4:I(MCM6^1)+I(MCM6^2)	0.0193
472	RFC4	HRASLS	RFC4:HRASLS	0.033
473	RFC4	NM.004702	RFC4:NM.004702	0.0397
474	CDCA7	MCM6	CDCA7:I(MCM6^1)	0.0044
475	CDCA7	MCM6	CDCA7:I(MCM6^1)+I(MCM6^2)	0.0044
476	CDCA7	Contig20217.RC	CDCA7:Contig20217.RC	0.0283
477	CDCA7	EGLN1	CDCA7:EGLN1	0.0242
478	CDCA7	NM.004702	CDCA7:NM.004702	0.0029
479	LOC643008	PITRM1	I(LOC643008^1):I(PITRM1^1)	0.0111
480	LOC643008	PITRM1	I(LOC643008^1)+I(LOC643008^2):I(PITRM1^1)+I(PITRM1^2)	0.0111
481	LOC643008	PITRM1	I(LOC643008^1)+I(LOC643008^2):I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3)	0.0111
482	LOC643008	Contig20217.RC	I(LOC643008^1):Contig20217.RC	0.0198
483	LOC643008	Contig20217.RC	I(LOC643008^1)+I(LOC643008^2):Contig20217.RC	0.0198
484	LOC643008	Contig20217.RC	I(LOC643008^1)+I(LOC643008^2)+I(LOC643008^3):Contig20217.RC	0.0198
485	LOC643008	C20orf46	I(LOC643008^1):I(C20orf46^1)	0.0454
486	LOC643008	C20orf46	I(LOC643008^1)+I(LOC643008^2):I(C20orf46^1)+I(C20orf46^2)	0.0454
487	LOC643008	C20orf46	I(LOC643008^1)+I(LOC643008^2)+I(LOC643008^3):I(C20orf46^1)	0.0454
488	MS4A7	LGP2	MS4A7:LGP2	0.0311
489	MCM6	PRC1	I(MCM6^1):PRC1	0.0093
490	MCM6	PRC1	I(MCM6^1)+I(MCM6^2):PRC1	0.0093

Appendix 9. Preselected Interactions from Deviance Test – NKI70 Data

Total Interactions = 508

#	V1	V2	Interactions	P-val[1]
491	MCM6	Contig20217.RC	I(MCM6^1):Contig20217.RC	0.0008
492	MCM6	Contig20217.RC	I(MCM6^1)+I(MCM6^2):Contig20217.RC	0.0008
493	MCM6	NM.004702	I(MCM6^1):NM.004702	0.0041
494	MCM6	NM.004702	I(MCM6^1)+I(MCM6^2):NM.004702	0.0041
495	IGFBP5	Contig20217.RC	IGFBP5:Contig20217.RC	0.0449
496	HRASLS	Contig20217.RC	HRASLS:Contig20217.RC	0.023
497	HRASLS	NM.004702	HRASLS:NM.004702	0.0086
498	PITRM1	Contig20217.RC	I(PITRM1^1):Contig20217.RC	0.009
499	PITRM1	Contig20217.RC	I(PITRM1^1)+I(PITRM1^2):Contig20217.RC	0.009
500	PITRM1	Contig20217.RC	I(PITRM1^1)+I(PITRM1^2)+I(PITRM1^3):Contig20217.RC	0.009
501	IGFBP5.1	Contig20217.RC	IGFBP5.1:Contig20217.RC	0.0268
502	NMU	PALM2.AKAP2	NMU:PALM2.AKAP2	0.0306
503	NMU	LGP2	NMU:LGP2	0.0103
504	NMU	PRC1	NMU:PRC1	0.0181
505	NMU	Contig20217.RC	NMU:Contig20217.RC	0.0218
506	LGP2	C20orf46	LGP2:I(C20orf46^1)	0.0427
507	LGP2	C20orf46	LGP2:I(C20orf46^1)+I(C20orf46^2)	0.0427
508	CENPA	NM.004702	CENPA:NM.004702	0.0108

NOTE:
I() is a special function to force the formula inside the parentheses to be executed first.

The P-values presented in the table are used for the purpose of screening interaction terms; it does not support any claims or conclusions. Additionally, there are a quite some cases when the interaction terms are non-estimable, therefore the p-values may stay the same for the interaction terms between particular pair of factors.

APPENDIX 10. All 3-degree Polynomial Terms, Including All Linear, Nonlinear and Interactions – NKI70 Data

Total Interactions = 735

Grade.Intermediate	Grade.Well	ER.Pos	N.GE4	Diam.GT2
ALDH4A1	AA555029.RC	NUSAP1	DIAPH3	TSPYL5
FLT1	EXT1	SCUBE2	C16orf61	BBC3
Contig35251.RC	RUNDC1	MMP9	OXCT1	GNAZ
GSTM3	AYTL2	SERF1A	GMPS	ECT2
UCHL5	RTN4RL1	ZNF533	RAB6B	GPR180
STK32B	COL4A2	Contig40831.RC	MTDH	PECI
PECI.1	SLC2A3	GPR126	FBXO31	DCK
C9orf30	AP2B1	MS4A7	CDCA7	RFC4
PALM2.AKAP2	NMU	IGFBP5.1	HRASLS	IGFBP5
EGLN1	CENPA	Contig20217.RC	PRC1	LGP2
Age^3	Age^2	Age	ESM1	NM.004702
FGF18	QSCN6L1^2	QSCN6L1	Contig63649.RC^2	Contig63649.RC
Contig32125.RC	DIAPH3.1^2	DIAPH3.1	FGF18^3	FGF18^2
RP5.860F19.3	DIAPH3.2^2	DIAPH3.2	Contig32125.RC^3	Contig32125.RC^2
WISP1^2	WISP1	KNTC2^2	KNTC2	RP5.860F19.3^2
TGFB3^2	TGFB3	CDC42BPA^2	CDC42BPA	WISP1^3
ORC6L	DTL^2	DTL	MELK^2	MELK
MCM6	LOC643008^3	LOC643008^2	LOC643008	ORC6L^2
C20orf46	PITRM1^3	PITRM1^2	PITRM1	MCM6^2
Age:Contig40831.RC	Age^3:GNAZ	Age^2:GNAZ	Age:GNAZ	C20orf46^2
Contig63649.RC:WISP1	Contig63649.RC^2:RUNDC1	Contig63649.RC:RUNDC1	Age^3:Contig40831.RC	Age^2:Contig40831.RC
Contig63649.RC^2:WISP1^3	Contig63649.RC^2:WISP1^2	Contig63649.RC^2:WISP1	Contig63649.RC:WISP1^3	Contig63649.RC:WISP1^2
DIAPH3:QSCN6L1	Contig63649.RC^2:CDC42BPA^2	Contig63649.RC^2:CDC42BPA	Contig63649.RC:CDC42BPA^2	Contig63649.RC:CDC42BPA
DIAPH3:UCHL5	DIAPH3:GPR180	DIAPH3:GSTM3	DIAPH3:GMPS	DIAPH3:QSCN6L1^2
DIAPH3:DTL^2	DIAPH3:DTL	DIAPH3:MELK^2	DIAPH3:MELK	DIAPH3:Contig40831.RC
DIAPH3:MCM6^2	DIAPH3:MCM6	DIAPH3:CDCA7	DIAPH3:ORC6L^2	DIAPH3:ORC6L
DIAPH3:CENPA	DIAPH3:PITRM1^3	DIAPH3:PITRM1^2	DIAPH3:PITRM1	DIAPH3:HRASLS
NUSAP1:TGFB3^2	NUSAP1:TGFB3	NUSAP1:RAB6B	NUSAP1:DIAPH3.2^2	NUSAP1:DIAPH3.2
NUSAP1:CENPA	NUSAP1:HRASLS	NUSAP1:MCM6^2	NUSAP1:MCM6	NUSAP1:CDCA7
ALDH4A1:DTL^2	ALDH4A1:DTL	ALDH4A1:RTN4RL1	AA555029.RC:RTN4RL1	AA555029.RC:FLT1
QSCN6L1^2:DIAPH3.1^2	QSCN6L1^2:DIAPH3.1	QSCN6L1:DIAPH3.1^2	QSCN6L1:DIAPH3.1	ALDH4A1:AP2B1
QSCN6L1:C16orf61	QSCN6L1^2:DIAPH3.2^2	QSCN6L1^2:DIAPH3.2	QSCN6L1:DIAPH3.2^2	QSCN6L1:DIAPH3.2
QSCN6L1:KNTC2^2	QSCN6L1:KNTC2	QSCN6L1^2:ECT2	QSCN6L1:ECT2	QSCN6L1^2:C16orf61
QSCN6L1:MTDH	QSCN6L1^2:SERF1A	QSCN6L1:SERF1A	QSCN6L1^2:KNTC2^2	QSCN6L1^2:KNTC2
QSCN6L1:TGFB3^2	QSCN6L1:TGFB3	QSCN6L1^2:Contig40831.RC	QSCN6L1:Contig40831.RC	QSCN6L1^2:MTDH

Appendix 10. All 3-degree Polynomial Terms, Including All Linear, Nonlinear and Interactions – NKI70 Data

QSCN6L1^2:MELK	QSCN6L1:MELK^2	QSCN6L1:MELK	QSCN6L1^2:TGFB3^2	QSCN6L1^2:TGFB3
QSCN6L1^2:DTL^2	QSCN6L1^2:DTL	QSCN6L1:DTL^2	QSCN6L1:DTL	QSCN6L1^2:MELK^2
QSCN6L1^2:ORC6L	QSCN6L1:ORC6L^2	QSCN6L1:ORC6L	QSCN6L1^2:FBXO31	QSCN6L1:FBXO31
QSCN6L1^2:CDCA7	QSCN6L1:CDCA7	QSCN6L1^2:RFC4	QSCN6L1:RFC4	QSCN6L1^2:ORC6L^2
QSCN6L1:Contig20217.RC	QSCN6L1^2:MCM6^2	QSCN6L1^2:MCM6	QSCN6L1:MCM6^2	QSCN6L1:MCM6
QSCN6L1^2:NM.004702	QSCN6L1:NM.004702	QSCN6L1^2:CENPA	QSCN6L1:CENPA	QSCN6L1^2:Contig20217.RC
FGF18:WISP1^2	FGF18:WISP1	FGF18^3:SCUBE2	FGF18^2:SCUBE2	FGF18:SCUBE2
FGF18^3:WISP1	FGF18^2:WISP1^3	FGF18^2:WISP1^2	FGF18^2:WISP1	FGF18:WISP1^3
FGF18^2:TGFB3	FGF18:TGFB3^2	FGF18:TGFB3	FGF18^3:WISP1^3	FGF18^3:WISP1^2
FGF18^2:STK32B	FGF18:STK32B	FGF18^3:TGFB3^2	FGF18^3:TGFB3	FGF18^2:TGFB3^2
FGF18:CDCA7	FGF18^3:DCK	FGF18^2:DCK	FGF18:DCK	FGF18^3:STK32B
FGF18^2:MCM6	FGF18:MCM6^2	FGF18:MCM6	FGF18^3:CDCA7	FGF18^2:CDCA7
FGF18:PITRM1^2	FGF18:PITRM1	FGF18^3:MCM6^2	FGF18^3:MCM6	FGF18^2:MCM6^2
FGF18^3:PITRM1	FGF18^2:PITRM1^3	FGF18^2:PITRM1^2	FGF18^2:PITRM1	FGF18:PITRM1^3
FGF18^3:NMU	FGF18^2:NMU	FGF18:NMU	FGF18^3:PITRM1^3	FGF18^3:PITRM1^2
DIAPH3.1:RUNDC1	DIAPH3.1^2:DIAPH3.2^2	DIAPH3.1^2:DIAPH3.2	DIAPH3.1:DIAPH3.2^2	DIAPH3.1:DIAPH3.2
DIAPH3.1^2:GMPS	DIAPH3.1:GMPS	DIAPH3.1^2:ECT2	DIAPH3.1:ECT2	DIAPH3.1^2:RUNDC1
DIAPH3.1:GSTM3	DIAPH3.1^2:KNTC2^2	DIAPH3.1^2:KNTC2	DIAPH3.1:KNTC2^2	DIAPH3.1:KNTC2
DIAPH3.1^2:Contig40831.RC	DIAPH3.1:Contig40831.RC	DIAPH3.1^2:UCHL5	DIAPH3.1:UCHL5	DIAPH3.1^2:GSTM3
DIAPH3.1:MELK	DIAPH3.1^2:TGFB3^2	DIAPH3.1^2:TGFB3	DIAPH3.1:TGFB3^2	DIAPH3.1:TGFB3
DIAPH3.1:DTL^2	DIAPH3.1:DTL	DIAPH3.1^2:MELK^2	DIAPH3.1^2:MELK	DIAPH3.1:MELK^2
DIAPH3.1:ORC6L	DIAPH3.1^2:GPR126	DIAPH3.1:GPR126	DIAPH3.1^2:DTL^2	DIAPH3.1^2:DTL
DIAPH3.1^2:RFC4	DIAPH3.1:RFC4	DIAPH3.1^2:ORC6L^2	DIAPH3.1^2:ORC6L	DIAPH3.1:ORC6L^2
DIAPH3.1^2:MCM6	DIAPH3.1:MCM6^2	DIAPH3.1:MCM6	DIAPH3.1^2:CDCA7	DIAPH3.1:CDCA7
DIAPH3.1^2:PITRM1	DIAPH3.1:PITRM1^3	DIAPH3.1:PITRM1^2	DIAPH3.1:PITRM1	DIAPH3.1^2:MCM6^2
DIAPH3.1:Contig20217.RC	DIAPH3.1^2:NMU	DIAPH3.1:NMU	DIAPH3.1^2:PITRM1^3	DIAPH3.1^2:PITRM1^2
Contig32125.RC^2:MTDH	Contig32125.RC:MTDH	DIAPH3.1^2:CENPA	DIAPH3.1:CENPA	DIAPH3.1^2:Contig20217.RC
Contig32125.RC:MCM6	Contig32125.RC^3:CDCA7	Contig32125.RC^2:CDCA7	Contig32125.RC:CDCA7	Contig32125.RC^3:MTDH
Contig32125.RC^3:MCM6^2	Contig32125.RC^3:MCM6	Contig32125.RC^2:MCM6^2	Contig32125.RC^2:MCM6	Contig32125.RC:MCM6^2
BBC3:MS4A7	BBC3:Contig40831.RC	BBC3:AYTL2	BBC3:KNTC2^2	BBC3:KNTC2
DIAPH3.2:C16orf61	BBC3:Contig20217.RC	BBC3:PALM2.AKAP2	BBC3:IGFBP5.1	BBC3:IGFBP5
DIAPH3.2^2:GMPS	DIAPH3.2:GMPS	DIAPH3.2^2:ECT2	DIAPH3.2:ECT2	DIAPH3.2^2:C16orf61
DIAPH3.2:UCHL5	DIAPH3.2^2:KNTC2^2	DIAPH3.2^2:KNTC2	DIAPH3.2:KNTC2^2	DIAPH3.2:KNTC2
DIAPH3.2:TGFB3^2	DIAPH3.2:TGFB3	DIAPH3.2^2:Contig40831.RC	DIAPH3.2:Contig40831.RC	DIAPH3.2^2:UCHL5
DIAPH3.2^2:MELK	DIAPH3.2:MELK^2	DIAPH3.2:MELK	DIAPH3.2^2:TGFB3^2	DIAPH3.2^2:TGFB3
DIAPH3.2^2:DTL^2	DIAPH3.2^2:DTL	DIAPH3.2:DTL^2	DIAPH3.2:DTL	DIAPH3.2^2:MELK^2

Appendix 10. All 3-degree Polynomial Terms, Including All Linear, Nonlinear and Interactions – NKI70 Data

Total Interactions = 735

DIAPH3.2:CDCA7	DIAPH3.2^2:ORC6L^2	DIAPH3.2^2:ORC6L	DIAPH3.2:ORC6L^2	DIAPH3.2:ORC6L
DIAPH3.2^2:MCM6^2	DIAPH3.2^2:MCM6	DIAPH3.2:MCM6^2	DIAPH3.2:MCM6	DIAPH3.2^2:CDCA7
DIAPH3.2:CENPA	DIAPH3.2^2:Contig20217.RC	DIAPH3.2:Contig20217.RC	DIAPH3.2^2:LGP2	DIAPH3.2:LGP2
RP5.860F19.3^2:Contig40831.RC	RP5.860F19.3:Contig40831.RC	DIAPH3.2^2:NM.004702	DIAPH3.2:NM.004702	DIAPH3.2^2:CENPA
RP5.860F19.3:RFC4	RP5.860F19.3^2:MELK^2	RP5.860F19.3^2:MELK	RP5.860F19.3:MELK^2	RP5.860F19.3:MELK
RP5.860F19.3^2:NM.004702	RP5.860F19.3:NM.004702	RP5.860F19.3^2:EGLN1	RP5.860F19.3:EGLN1	RP5.860F19.3^2:RFC4
C16orf61:MELK^2	C16orf61:MELK	C16orf61:TGFB3^2	C16orf61:TGFB3	C16orf61:GMPS
C16orf61:MCM6^2	C16orf61:MCM6	C16orf61:CDCA7	C16orf61:DTL^2	C16orf61:DTL
SCUBE2:PECI	SCUBE2:RAB6B	SCUBE2:AYTL2	SCUBE2:Contig35251.RC	C16orf61:HRASLS
SCUBE2:LOC643008	SCUBE2:PECI.1	SCUBE2:STK32B	SCUBE2:TGFB3^2	SCUBE2:TGFB3
FLT1:Contig35251.RC	FLT1:RUNDC1	FLT1:MMP9	SCUBE2:LOC643008^3	SCUBE2:LOC643008^2
FLT1:EGLN1	FLT1:AP2B1	FLT1:WISP1^3	FLT1:WISP1^2	FLT1:WISP1
OXCT1:NMU	OXCT1:IGFBP5.1	GNAZ:LGP2	GNAZ:MS4A7	GNAZ:CDCA7
MMP9:IGFBP5.1	MMP9:IGFBP5	MMP9:COL4A2	MMP9:CDC42BPA^2	MMP9:CDC42BPA
RUNDC1:LGP2	RUNDC1:TGFB3^2	RUNDC1:TGFB3	RUNDC1:ECT2	RUNDC1:Contig35251.RC
Contig35251.RC:CDC42BPA^2	Contig35251.RC:CDC42BPA	Contig35251.RC:KNTC2^2	Contig35251.RC:KNTC2	Contig35251.RC:ECT2
Contig35251.RC:TGFB3	Contig35251.RC:Contig40831.RC	Contig35251.RC:MTDH	Contig35251.RC:ZNF533	Contig35251.RC:RAB6B
Contig35251.RC:NM.004702	Contig35251.RC:CENPA	Contig35251.RC:ORC6L^2	Contig35251.RC:ORC6L	Contig35251.RC:TGFB3^2
ECT2:TGFB3	ECT2:Contig40831.RC	ECT2:UCHL5	ECT2:KNTC2^2	ECT2:KNTC2
ECT2:DTL^2	ECT2:DTL	ECT2:MELK^2	ECT2:MELK	ECT2:TGFB3^2
ECT2:MCM6^2	ECT2:MCM6	ECT2:CDCA7	ECT2:GPR126	ECT2:DCK
GMPS:DTL	GMPS:KNTC2^2	GMPS:KNTC2	ECT2:CENPA	ECT2:HRASLS
GMPS:PRC1	GMPS:MCM6^2	GMPS:MCM6	GMPS:CDCA7	GMPS:DTL^2
KNTC2:Contig40831.RC	KNTC2^2:RAB6B	KNTC2:RAB6B	GMPS:EGLN1	GMPS:Contig20217.RC
KNTC2^2:TGFB3^2	KNTC2^2:TGFB3	KNTC2:TGFB3^2	KNTC2:TGFB3	KNTC2^2:Contig40831.RC
KNTC2:GPR126	KNTC2^2:MELK^2	KNTC2^2:MELK	KNTC2:MELK^2	KNTC2:MELK
KNTC2^2:CDCA7	KNTC2:CDCA7	KNTC2^2:PECI.1	KNTC2:PECI.1	KNTC2^2:GPR126
KNTC2^2:LOC643008^2	KNTC2^2:LOC643008	KNTC2:LOC643008^3	KNTC2:LOC643008^2	KNTC2:LOC643008
KNTC2^2:IGFBP5.1	KNTC2:IGFBP5.1	KNTC2^2:IGFBP5	KNTC2:IGFBP5	KNTC2^2:LOC643008^3
WISP1:UCHL5	KNTC2^2:NM.004702	KNTC2:NM.004702	KNTC2^2:Contig20217.RC	KNTC2:Contig20217.RC
WISP1^2:DTL	WISP1:DTL^2	WISP1:DTL	WISP1^3:UCHL5	WISP1^2:UCHL5
WISP1^2:STK32B	WISP1:STK32B	WISP1^3:DTL^2	WISP1^3:DTL	WISP1^2:DTL^2
WISP1^2:ORC6L^2	WISP1^2:ORC6L	WISP1:ORC6L^2	WISP1:ORC6L	WISP1^3:STK32B
WISP1^3:CDCA7	WISP1^2:CDCA7	WISP1:CDCA7	WISP1^3:ORC6L^2	WISP1^3:ORC6L
WISP1^3:MCM6	WISP1^2:MCM6^2	WISP1^2:MCM6	WISP1:MCM6^2	WISP1:MCM6
WISP1^2:PITRM1	WISP1:PITRM1^3	WISP1:PITRM1^2	WISP1:PITRM1	WISP1^3:MCM6^2

Appendix 10. All 3-degree Polynomial Terms, Including All Linear, Nonlinear and Interactions – NKI70 Data

Total Interactions = 735

WISP1^3:PITRM1^3	WISP1^3:PITRM1^2	WISP1^3:PITRM1	WISP1^2:PITRM1^3	WISP1^2:PITRM1^2
WISP1^2:ESM1	WISP1:ESM1	WISP1^3:PRC1	WISP1^2:PRC1	WISP1:PRC1
CDC42BPA^2:LOC643008	CDC42BPA:LOC643008^3	CDC42BPA:LOC643008^2	CDC42BPA:LOC643008	WISP1^3:ESM1
SERF1A:MCM6^2	SERF1A:MCM6	SERF1A:CDCA7	CDC42BPA^2:LOC643008^3	CDC42BPA^2:LOC643008^2
AYTL2:C9orf30	AYTL2:MS4A7	AYTL2:STK32B	AYTL2:RTN4RL1	SERF1A:NMU
GPR180:MELK^2	GPR180:MELK	GSTM3:LGP2	GSTM3:IGFBP5.1	AYTL2:NMU
RAB6B:RFC4	RAB6B:ORC6L^2	RAB6B:ORC6L	RAB6B:Contig40831.RC	RAB6B:UCHL5
RAB6B:EGLN1	RAB6B:Contig20217.RC	RAB6B:PRC1	RAB6B:C9orf30	RAB6B:CDCA7
UCHL5:TGFB3^2	UCHL5:TGFB3	UCHL5:Contig40831.RC	RTN4RL1:PALM2.AKAP2	RTN4RL1:HRASLS
MTDH:PITRM1^2	MTDH:PITRM1	MTDH:CDCA7	PECI:PALM2.AKAP2	UCHL5:MS4A7
Contig40831.RC:MELK	Contig40831.RC:TGFB3^2	Contig40831.RC:TGFB3	MTDH:EGLN1	MTDH:PITRM1^3
Contig40831.RC:CDCA7	Contig40831.RC:ORC6L^2	Contig40831.RC:ORC6L	Contig40831.RC:STK32B	Contig40831.RC:MELK^2
Contig40831.RC:PITRM1^3	Contig40831.RC:PITRM1^2	Contig40831.RC:PITRM1	Contig40831.RC:MCM6^2	Contig40831.RC:MCM6
TGFB3^2:MELK	TGFB3:MELK^2	TGFB3:MELK	Contig40831.RC:NM.004702	Contig40831.RC:CENPA
TGFB3:ORC6L^2	TGFB3:ORC6L	TGFB3^2:DCK	TGFB3:DCK	TGFB3^2:MELK^2
TGFB3:CDCA7	TGFB3^2:RFC4	TGFB3:RFC4	TGFB3^2:ORC6L^2	TGFB3^2:ORC6L
TGFB3^2:MCM6^2	TGFB3^2:MCM6	TGFB3:MCM6^2	TGFB3:MCM6	TGFB3^2:CDCA7
TGFB3:PITRM1^3	TGFB3:PITRM1^2	TGFB3:PITRM1	TGFB3^2:AP2B1	TGFB3:AP2B1
TGFB3^2:IGFBP5.1	TGFB3:IGFBP5.1	TGFB3^2:PITRM1^3	TGFB3^2:PITRM1^2	TGFB3^2:PITRM1
MELK:RFC4	TGFB3^2:NM.004702	TGFB3:NM.004702	TGFB3^2:Contig20217.RC	TGFB3:Contig20217.RC
MELK:MCM6^2	MELK:MCM6	MELK^2:CDCA7	MELK:CDCA7	MELK^2:RFC4
MELK:PITRM1^3	MELK:PITRM1^2	MELK:PITRM1	MELK^2:MCM6^2	MELK^2:MCM6
MELK^2:CENPA	MELK:CENPA	MELK^2:PITRM1^3	MELK^2:PITRM1^2	MELK^2:PITRM1
DTL^2:MCM6	DTL:MCM6^2	DTL:MCM6	COL4A2:LGP2	COL4A2:HRASLS
STK32B:PRC1	STK32B:FBXO31	DTL^2:NMU	DTL:NMU	DTL^2:MCM6^2
FBXO31:PITRM1	FBXO31:HRASLS	DCK:NM.004702	DCK:SLC2A3	DCK:GPR126
PECI.1:RFC4	SLC2A3:LGP2	SLC2A3:CDCA7	FBXO31:PITRM1^3	FBXO31:PITRM1^2
ORC6L^2:LOC643008	ORC6L:LOC643008^3	ORC6L:LOC643008^2	ORC6L:LOC643008	PECI.1:PALM2.AKAP2
ORC6L:CENPA	ORC6L^2:Contig20217.RC	ORC6L:Contig20217.RC	ORC6L^2:LOC643008^3	ORC6L^2:LOC643008^2
RFC4:HRASLS	RFC4:MCM6^2	RFC4:MCM6	RFC4:CDCA7	ORC6L^2:CENPA
CDCA7:EGLN1	CDCA7:Contig20217.RC	CDCA7:MCM6^2	CDCA7:MCM6	RFC4:NM.004702
LOC643008^2:PITRM1	LOC643008:PITRM1^3	LOC643008:PITRM1^2	LOC643008:PITRM1	CDCA7:NM.004702
LOC643008^3:PITRM1^3	LOC643008^3:PITRM1^2	LOC643008^3:PITRM1	LOC643008^2:PITRM1^3	LOC643008^2:PITRM1^2
LOC643008:C20orf46^2	LOC643008:C20orf46	LOC643008^3:Contig20217.RC	LOC643008^2:Contig20217.RC	LOC643008:Contig20217.RC
MS4A7:LGP2	LOC643008^3:C20orf46^2	LOC643008^3:C20orf46	LOC643008^2:C20orf46^2	LOC643008^2:C20orf46
MCM6:NM.004702	MCM6^2:Contig20217.RC	MCM6:Contig20217.RC	MCM6^2:PRC1	MCM6:PRC1

Appendix 10. All 3-degree Polynomial Terms, Including All Linear, Nonlinear and Interactions – NKI70 Data

Total Interactions = 735

PITRM1:Contig20217.RC	HRASLS:NM.004702	HRASLS:Contig20217.RC	IGFBP5:Contig20217.RC	MCM6^2:NM.004702
NMU:LGP2	NMU:PALM2.AKAP2	IGFBP5.1:Contig20217.RC	PITRM1^3:Contig20217.RC	PITRM1^2:Contig20217.RC
CENPA:NM.004702	LGP2:C20orf46^2	LGP2:C20orf46	NMU:Contig20217.RC	NMU:PRC1

APPENDIX 11. Biased Estimates of the Regression Coefficients from Lasso Interaction Model – NKI70 Data

	Coef	HR	SE (Coef)	z	Pr(>|z|)
Age:TSPYL5	-0.0068	0.993	0.0176	-0.39	0.6991
Age:Contig63649.RC	0.0089	1.009	0.0246	0.36	0.7165
Age:QSCN6L1	0.0064	1.006	0.0382	0.17	0.8676
Age:Contig32125.RC	0.0213	1.022	0.0319	0.67	0.5047
Age:MMP9	0.0106	1.011	0.0243	0.44	0.6612
Age:RUNDC1	0.0359	1.037	0.0304	1.18	0.2383
Age:KNTC2	-0.0526	0.949	0.0383	-1.37	0.1701
Age:RAB6B	-0.0034	0.997	0.0185	-0.18	0.8534
Age:ZNF533	-0.0119	0.988	0.0137	-0.87	0.3864
Age:COL4A2	0.0123	1.012	0.0348	0.35	0.7236
Age:GPR126	-0.0397	0.961	0.0219	-1.82	0.0695
Age:PECI.1	-0.0156	0.984	0.0338	-0.46	0.6442
Age:CDCA7	0.0135	1.014	0.0229	0.59	0.5544
Age:LOC643008	-0.0013	0.999	0.0210	-0.06	0.9490
Age:MS4A7	-0.0383	0.962	0.0242	-1.58	0.1134
Age:IGFBP5	0.0387	1.039	0.0152	2.54	0.0110
Age:HRASLS	-0.0377	0.963	0.0245	-1.54	0.1242
Age:PRC1	0.0313	1.032	0.0404	0.77	0.4384
Age:CENPA	0.0047	1.005	0.0318	0.15	0.8815
Age:NM.004702	0.0440	1.045	0.0313	1.40	0.1605
Age:ESM1	0.0208	1.021	0.0243	0.85	0.3928
Age:C20orf46	-0.0055	0.994	0.0235	-0.24	0.8140
Age:N.GE4	0.0107	1.011	0.0116	0.92	0.3562
Age:ER.Pos	-0.0289	0.972	0.0172	-1.68	0.0937
Age:Grade.Well	-0.0003	1.000	0.0141	-0.02	0.9812

APPENDIX 12. Unbiased Regression Coefficients Corresponding to the 25 Covariate Terms as Retained by Lasso Interaction Model – NKI70 Data

	Coef	HR	SE (Coef)	z	Pr(>|z|)
Age:TSPYL5	-0.0152	0.985	0.0245	-0.62	0.5342
Age:Contig63649.RC	0.0176	1.018	0.0298	0.59	0.5543
Age:QSCN6L1	-0.0157	0.984	0.0556	-0.28	0.7777
Age:Contig32125.RC	0.0708	1.073	0.0460	1.54	0.1238
Age:MMP9	0.0331	1.034	0.0347	0.95	0.3398
Age:RUNDC1	0.1172	1.124	0.0400	2.93	0.0034
Age:KNTC2	-0.2158	0.806	0.0743	-2.90	0.0037
Age:RAB6B	-0.0252	0.975	0.0250	-1.01	0.3142
Age:ZNF533	0.0007	1.001	0.0187	0.04	0.9691
Age:COL4A2	0.0834	1.087	0.0446	1.87	0.0617
Age:GPR126	-0.1001	0.905	0.0339	-2.95	0.0032
Age:PECI.1	-0.0884	0.915	0.0496	-1.78	0.0749
Age:CDCA7	0.0422	1.043	0.0355	1.19	0.2340
Age:LOC643008	-0.0243	0.976	0.0301	-0.81	0.4198
Age:MS4A7	-0.0408	0.960	0.0311	-1.31	0.1897
Age:IGFBP5	0.0961	1.101	0.0274	3.51	0.0005
Age:HRASLS	-0.0780	0.925	0.0476	-1.64	0.1016
Age:PRC1	0.0994	1.105	0.0520	1.91	0.0560
Age:CENPA	0.0499	1.051	0.0467	1.07	0.2852
Age:NM.004702	0.1145	1.121	0.0402	2.85	0.0044
Age:ESM1	0.0826	1.086	0.0390	2.12	0.0340
Age:C20orf46	-0.0062	0.994	0.0410	-0.15	0.8802
Age:N.GE4	0.0106	1.011	0.0146	0.72	0.4692
Age:ER.Pos	-0.0575	0.944	0.0258	-2.23	0.0260
Age:Grade.Well	0.0051	1.005	0.0177	0.29	0.7735

APPENDIX 13. Biased estimates of Regression Coefficients from Ridge Cox Linear Model – NKI70 Data

	Coef	HR	SE (Coef)	z	Pr (>|z|)
Diam.GT2	0.3059	1.358	1.0199	0.30	0.7642
N.GE4	0.2979	1.347	1.0546	0.28	0.7776
ER.Pos	-0.3960	0.673	2.4597	-0.16	0.8721
Grade.Well	-0.1101	0.896	1.3102	-0.08	0.9331
Grade.Intermediate	0.0664	1.069	1.0708	0.06	0.9505
Age	-0.0472	0.954	0.0904	-0.52	0.6017
TSPYL5	-0.1953	0.823	1.3696	-0.14	0.8866
Contig63649.RC	0.1990	1.220	2.2566	0.09	0.9297
DIAPH3	0.0101	1.010	3.8526	0.00	0.9979
NUSAP1	0.3281	1.388	4.0716	0.08	0.9358
AA555029.RC	-0.0435	0.957	3.0798	-0.01	0.9887
ALDH4A1	0.1015	1.107	3.6532	0.03	0.9778
QSCN6L1	0.3323	1.394	3.5595	0.09	0.9256
FGF18	-0.0031	0.997	2.2191	0.00	0.9989
DIAPH3.1	-0.0805	0.923	5.3835	-0.03	0.9881
Contig32125.RC	0.3070	1.359	2.7924	0.12	0.9125
BBC3	0.0126	1.013	4.2787	0.00	0.9976
DIAPH3.2	0.0178	1.018	7.6300	0.00	0.9981
RP5.860F19.3	0.0888	1.093	2.6084	0.03	0.9728
C16orf61	0.0285	1.029	4.9999	0.02	0.9955
SCUBE2	-0.0908	0.913	1.3154	-0.07	0.9450
EXT1	0.0620	1.064	3.9930	0.02	0.9876
FLT1	0.2075	1.231	3.4875	0.06	0.9526
GNAZ	0.1627	1.177	2.7607	0.06	0.9530
OXCT1	0.1719	1.188	4.0154	0.04	0.9658
MMP9	0.2197	1.246	1.7364	0.13	0.8993
RUNDC1	0.3551	1.426	3.1603	0.11	0.9105
Contig35251.RC	0.0607	1.063	3.3179	0.02	0.9854
ECT2	0.2324	1.262	3.7434	0.06	0.9505
GMPS	0.0054	1.005	3.3384	0.00	0.9987
KNTC2	-0.3248	0.723	3.4907	-0.09	0.9259
WISP1	-0.0225	0.978	3.7847	-0.01	0.9953
CDC42BPA	0.0013	1.001	5.1306	0.00	0.9998
SERF1A	-0.1110	0.895	4.5575	-0.02	0.9806
AYTL2	0.0405	1.041	3.7264	0.01	0.9913
GSTM3	0.0779	1.081	1.6793	0.05	0.9630
GPR180	-0.0856	0.918	3.4455	-0.02	0.9802
RAB6B	-0.1327	0.876	1.3581	-0.10	0.9221
ZNF533	-0.5648	0.568	1.2358	-0.46	0.6476
RTN4RL1	-0.1358	0.873	3.1103	-0.04	0.9652
UCHL5	-0.1574	0.854	4.5542	-0.03	0.9724
PECI	-0.1985	0.820	4.8088	-0.04	0.9671
MTDH	0.0283	1.029	4.4042	0.01	0.9949
Contig40831.RC	0.0729	1.076	3.8774	0.02	0.9850

	Coef	HR	SE (Coef)	z	Pr (>|z|)
TGFB3	-0.1501	0.861	3.1853	-0.05	0.9624
MELK	-0.0168	0.983	5.0699	0.00	0.9974
COL4A2	0.2496	1.284	3.7094	0.07	0.9463
DTL	0.1159	1.123	4.8575	0.02	0.9810
STK32B	-0.1375	0.872	4.0412	-0.03	0.9729
DCK	-0.0068	0.993	3.5546	0.00	0.9985
FBXO31	-0.0067	0.993	4.0274	0.00	0.9987
GPR126	-0.4866	0.615	1.9416	-0.25	0.8021
SLC2A3	-0.1399	0.869	2.7822	-0.05	0.9599
PECI.1	-0.2692	0.764	5.3297	-0.05	0.9597
ORC6L	0.1949	1.215	2.9478	0.07	0.9473
RFC4	0.0724	1.075	4.5109	0.02	0.9872
CDCA7	0.1341	1.144	1.8760	0.07	0.9430
LOC643008	-0.1801	0.835	1.8487	-0.10	0.9224
MS4A7	-0.5516	0.576	1.9365	-0.28	0.7757
MCM6	-0.0832	0.920	4.7670	-0.02	0.9861
AP2B1	-0.0092	0.991	3.6808	0.00	0.9980
C9orf30	0.0880	1.092	5.3224	0.02	0.9868
IGFBP5	0.4172	1.518	5.8902	0.07	0.9435
HRASLS	-0.3975	0.672	2.4114	-0.16	0.8691
PITRM1	-0.2750	0.760	3.8104	-0.07	0.9425
IGFBP5.1	0.3604	1.434	6.1854	0.06	0.9535
NMU	-0.0588	0.943	1.5366	-0.04	0.9695
PALM2.AKAP2	-0.1418	0.868	4.2673	-0.03	0.9735
LGP2	0.1985	1.220	2.6075	0.08	0.9393
PRC1	0.3542	1.425	3.4851	0.10	0.9190
Contig20217.RC	-0.2426	0.785	3.5542	-0.07	0.9456
CENPA	0.3056	1.357	2.4890	0.12	0.9023
EGLN1	-0.2315	0.793	4.2708	-0.05	0.9568
NM.004702	0.3567	1.429	3.9325	0.09	0.9277
ESM1	0.2628	1.301	1.8927	0.14	0.8896
C20orf46	-0.1501	0.861	2.0366	-0.07	0.9412

APPENDIX 14. Biased estimates of the Regression Coefficients from Elastic-Net Cox Linear Model – NKI70 Data

	Coef	HR	SE (coef)	z	Pr(>|z|)
Diam.GT2	0.0840	1.0876	0.8061	0.10	0.9170
N.GE4	0.1531	1.1654	0.8841	0.17	0.8625
ER.Pos	-0.1239	0.8835	1.6094	-0.08	0.9386
Grade.Well	-0.0468	0.9543	1.1061	-0.04	0.9663
Grade.Intermediate	0.0060	1.0060	0.8958	0.01	0.9946
Age	-0.0090	0.9910	0.0707	-0.13	0.8983
TSPYL5	-0.0665	0.9357	1.1965	-0.06	0.9557
Contig63649.RC	0.1670	1.1818	1.8354	0.09	0.9275
DIAPH3	0.0144	1.0145	3.0319	0.00	0.9962
NUSAP1	0.2179	1.2434	3.0069	0.07	0.9422
AA555029.RC	-0.0045	0.9956	2.5999	0.00	0.9986
ALDH4A1	0.0629	1.0650	3.2493	0.02	0.9845
QSCN6L1	0.3011	1.3513	2.9176	0.10	0.9178
Contig32125.RC	0.3228	1.3809	2.0882	0.15	0.8772
DIAPH3.2	0.0523	1.0537	5.7656	0.01	0.9928
C16orf61	0.0454	1.0464	4.1133	0.01	0.9912
SCUBE2	-0.0307	0.9697	1.1255	-0.03	0.9782
EXT1	0.1491	1.1608	3.0858	0.05	0.9615
FLT1	0.2421	1.2740	2.8261	0.09	0.9317
GNAZ	0.1310	1.1400	2.1298	0.06	0.9509
OXCT1	0.2318	1.2609	3.2823	0.07	0.9437
MMP9	0.1094	1.1156	1.3752	0.08	0.9366
RUNDC1	0.3008	1.3509	2.6440	0.11	0.9094
Contig35251.RC	0.0629	1.0649	2.6053	0.02	0.9808
ECT2	0.2391	1.2702	2.9278	0.08	0.9349
GMPS	0.0302	1.0306	2.7752	0.01	0.9913
KNTC2	-0.1962	0.8218	3.0548	-0.06	0.9488
WISP1	-0.0090	0.9910	2.6321	0.00	0.9973
CDC42BPA	-0.0019	0.9981	3.4245	0.00	0.9996
SERF1A	-0.1465	0.8637	3.7212	-0.04	0.9686
AYTL2	0.0140	1.0141	3.0942	0.00	0.9964
GPR180	-0.0494	0.9518	3.0623	-0.02	0.9871
ZNF533	-0.1914	0.8258	1.0059	-0.19	0.8491
RTN4RL1	-0.2318	0.7931	2.2914	-0.10	0.9194
UCHL5	-0.2571	0.7733	4.1538	-0.06	0.9506
PECI	-0.2238	0.7995	4.1992	-0.05	0.9575
MTDH	-0.0159	0.9842	3.6851	0.00	0.9966
Contig40831.RC	0.0171	1.0172	2.8525	0.01	0.9952
TGFB3	-0.1090	0.8967	2.5451	-0.04	0.9658
MELK	0.0136	1.0137	4.0957	0.00	0.9973
COL4A2	0.3612	1.4350	3.0329	0.12	0.9052
DTL	0.1611	1.1748	3.8112	0.04	0.9663
STK32B	-0.4084	0.6647	3.2324	-0.13	0.8995
DCK	-0.0338	0.9668	3.1372	-0.01	0.9914

	Coef	HR	SE (coef)	z	Pr(>|z|)
FBXO31	0.0136	1.0137	3.4892	0.00	0.9969
GPR126	-0.2281	0.7961	1.6410	-0.14	0.8895
SLC2A3	-0.1490	0.8615	2.1979	-0.07	0.9459
PECI.1	-0.3259	0.7219	4.3165	-0.08	0.9398
ORC6L	0.1507	1.1627	2.5465	0.06	0.9528
RFC4	0.1232	1.1311	3.9352	0.03	0.9750
CDCA7	0.0469	1.0480	1.6695	0.03	0.9776
LOC643008	-0.0141	0.9860	1.4676	-0.01	0.9923
MS4A7	-0.3220	0.7247	1.4817	-0.22	0.8280
C9orf30	0.2048	1.2273	3.9428	0.05	0.9586
IGFBP5	0.1950	1.2153	4.9570	0.04	0.9686
HRASLS	-0.2605	0.7706	2.0378	-0.13	0.8983
PITRM1	-0.1711	0.8428	3.2465	-0.05	0.9580
IGFBP5.1	0.2235	1.2504	5.3445	0.04	0.9666
NMU	-0.0048	0.9952	1.1286	0.00	0.9966
PALM2.AKAP2	-0.0959	0.9086	3.1001	-0.03	0.9753
LGP2	0.1652	1.1796	2.0051	0.08	0.9343
PRC1	0.2537	1.2887	2.8290	0.09	0.9286
Contig20217.RC	-0.2111	0.8097	2.3974	-0.09	0.9298
CENPA	0.1327	1.1419	2.1367	0.06	0.9505
EGLN1	-0.3252	0.7224	3.3604	-0.10	0.9229
NM.004702	0.1876	1.2064	3.0611	0.06	0.9511
ESM1	0.1222	1.1300	1.5323	0.08	0.9364
C20orf46	-0.1032	0.9020	1.4690	-0.07	0.9440

APPENDIX 15. Biased estimates of the Regression Coefficients from Elastic-Net Cox Polynomial Model – NKI70 Data

	Coef	HR	SE (coef)	z	Pr(>|z|)
Diam.GT2	0.0419	1.0428	24.3532	0.0017	0.9986
N.GE4	0.0957	1.1005	15.1464	0.0063	0.9950
ER.Pos	-0.1041	0.9011	11.3852	-0.0091	0.9927
Grade.Well	-0.0070	0.9930	5.1677	-0.0014	0.9989
NUSAP1	0.0405	1.0413	84.3027	0.0005	0.9996
SCUBE2	-0.0085	0.9916	3.8881	-0.0022	0.9983
EXT1	0.0215	1.0218	22.3565	0.0010	0.9992
FLT1	0.0130	1.0131	13.4149	0.0010	0.9992
OXCT1	0.0271	1.0275	46.6916	0.0006	0.9995
ECT2	0.0467	1.0478	8.6073	0.0054	0.9957
ZNF533	-0.1256	0.8820	4.5734	-0.0275	0.9781
RTN4RL1	-0.1681	0.8453	41.1847	-0.0041	0.9967
PECI	-0.1258	0.8818	39.1872	-0.0032	0.9974
COL4A2	0.2060	1.2287	44.7465	0.0046	0.9963
STK32B	-0.1571	0.8546	27.9478	-0.0056	0.9955
GPR126	-0.0451	0.9559	31.3893	-0.0014	0.9989
SLC2A3	-0.0127	0.9874	53.3869	-0.0002	0.9998
PECI.1	-0.2381	0.7881	24.5508	-0.0097	0.9923
RFC4	0.0184	1.0185	22.4870	0.0008	0.9993
CDCA7	0.0044	1.0045	14.5616	0.0003	0.9998
MS4A7	-0.1788	0.8363	16.5857	-0.0108	0.9914
IGFBP5	0.0730	1.0757	31.1613	0.0023	0.9981
IGFBP5.1	0.0903	1.0945	24.9787	0.0036	0.9971
PRC1	0.1276	1.1361	33.1341	0.0039	0.9969
CENPA	0.0664	1.0686	13.8298	0.0048	0.9962
EGLN1	-0.1042	0.9010	19.2744	-0.0054	0.9957
NM.004702	0.0927	1.0972	30.9679	0.0030	0.9976
Age	-0.0021	0.9979	115.3825	0.0000	1.0000
I(Age^2)	0.0000	1.0000	2.8618	0.0000	1.0000
I(Age^3)	0.0000	1.0000	0.0231	0.0000	1.0000
Contig63649.RC	0.0321	1.0326	45.1516	0.0007	0.9994
I(Contig63649.RC^2)	0.4989	1.6469	113.7782	0.0044	0.9965
QSCN6L1	0.1870	1.2056	102.1358	0.0018	0.9985
I(RP5.860F19.3^2)	-0.0089	0.9912	109.5608	-0.0001	0.9999
I(CDC42BPA^2)	-1.2436	0.2884	74.2103	-0.0168	0.9866
TGFB3	-0.0677	0.9346	9.2450	-0.0073	0.9942
ORC6L	0.0365	1.0372	11.6879	0.0031	0.9975
I(C20orf46^2)	-0.1326	0.8759	13.3133	-0.0100	0.9921
I(Contig63649.RC):I(RUNDC1)	-1.0414	0.3530	95.0696	-0.0110	0.9913
I(Contig63649.RC^2):I(RUNDC1)	-0.0120	0.9880	309.3306	0.0000	1.0000
I(Contig63649.RC):I(WISP1)	0.4958	1.6418	112.0984	0.0044	0.9965
I(Contig63649.RC):I(WISP1^2)	0.7943	2.2130	283.2898	0.0028	0.9978
I(Contig63649.RC):I(WISP1^3)	4.4659	86.9965	1437.0903	0.0031	0.9975

	Coef	HR	SE (coef)	z	Pr(>|z|)
I(Contig63649.RC^2):I(WISP1)	0.2551	1.2906	2027.2687	0.0001	0.9999
I(Contig63649.RC^2):I(WISP1^2)	2.5465	12.7620	2220.0226	0.0011	0.9991
I(Contig63649.RC^2):I(WISP1^3)	8.3787	4353.5237	17740.4986	0.0005	0.9996
I(Contig63649.RC):I(CDC42BPA)	0.0767	1.0797	104.6008	0.0007	0.9994
I(Contig63649.RC^2):I(CDC42BPA^2)	-2.5632	0.0771	7189.4251	-0.0004	0.9997
I(DIAPH3):I(MCM6)	-0.1804	0.8350	98.2633	-0.0018	0.9985
I(NUSAP1):I(RAB6B)	-0.1243	0.8831	70.8710	-0.0018	0.9986
I(NUSAP1):I(TGFB3)	0.0626	1.0646	39.0425	0.0016	0.9987
I(NUSAP1):I(CDCA7)	-0.0531	0.9482	135.4924	-0.0004	0.9997
I(MCM6):I(NUSAP1)	-0.0341	0.9664	131.0418	-0.0003	0.9998
I(NUSAP1):I(HRASLS)	-0.1452	0.8648	109.2030	-0.0013	0.9989
I(NUSAP1):I(CENPA)	-0.0184	0.9818	94.5703	-0.0002	0.9998
I(AA555029.RC):I(FLT1)	1.0221	2.7789	290.6267	0.0035	0.9972
I(AA555029.RC):I(RTN4RL1)	0.9512	2.5887	259.5234	0.0037	0.9971
I(RTN4RL1):I(ALDH4A1)	-1.4300	0.2393	138.8473	-0.0103	0.9918
I(ALDH4A1):I(DTL)	-0.4106	0.6632	76.4673	-0.0054	0.9957
I(ALDH4A1):I(AP2B1)	-0.8431	0.4304	311.2000	-0.0027	0.9978
I(QSCN6L1):I(C16orf61)	-0.7848	0.4562	251.6089	-0.0031	0.9975
I(QSCN6L1):I(KNTC2)	-0.0687	0.9336	272.2287	-0.0003	0.9998
I(QSCN6L1):I(SERF1A)	-0.4612	0.6305	159.3749	-0.0029	0.9977
I(SERF1A):I(QSCN6L1^2)	-0.0741	0.9286	725.7190	-0.0001	0.9999
I(QSCN6L1):I(MTDH)	-1.1332	0.3220	275.8762	-0.0041	0.9967
I(QSCN6L1^2):I(MTDH)	-0.0790	0.9241	854.4795	-0.0001	0.9999
I(QSCN6L1):I(Contig40831.RC)	-0.7485	0.4731	139.1295	-0.0054	0.9957
I(QSCN6L1):I(TGFB3^2)	0.6257	1.8696	180.5756	0.0035	0.9972
I(QSCN6L1):I(MELK)	-0.0039	0.9961	406.9697	0.0000	1.0000
I(QSCN6L1):I(ORC6L)	-0.3408	0.7112	110.9966	-0.0031	0.9976
I(CDCA7):I(QSCN6L1)	-0.1740	0.8403	191.4352	-0.0009	0.9993
I(QSCN6L1):I(Contig20217.RC)	-0.5274	0.5901	314.0915	-0.0017	0.9987
I(CENPA):I(QSCN6L1^2)	1.3803	3.9760	867.6301	0.0016	0.9987
I(QSCN6L1):I(NM.004702)	-0.9431	0.3894	224.8059	-0.0042	0.9967
I(FGF18^3):I(SCUBE2)	-0.1723	0.8417	119.0617	-0.0014	0.9988
I(WISP1):I(FGF18)	-0.1990	0.8195	101.8365	-0.0020	0.9984
I(TGFB3):I(FGF18^2)	-0.3834	0.6816	283.5840	-0.0014	0.9989
I(TGFB3):I(FGF18^3)	-0.0088	0.9912	273.3668	0.0000	1.0000
I(FGF18^2):I(STK32B)	-1.5689	0.2083	497.6667	-0.0032	0.9975
I(FGF18^3):I(STK32B)	-0.7603	0.4675	554.6285	-0.0014	0.9989
I(FGF18):I(DCK)	-0.4431	0.6420	409.0938	-0.0011	0.9991
I(FGF18^3):I(DCK)	-1.9970	0.1357	1912.3392	-0.0010	0.9992
I(MCM6):I(FGF18^3)	0.8442	2.3260	1243.9368	0.0007	0.9995
I(FGF18):I(PITRM1)	0.4207	1.5231	75.9386	0.0055	0.9956
I(FGF18):I(PITRM1^3)	1.7426	5.7121	391.5484	0.0045	0.9964
I(FGF18^3):I(PITRM1^3)	2.3801	10.8059	7439.1157	0.0003	0.9997
I(Contig40831.RC):I(DIAPH3.1)	-0.0662	0.9359	83.4397	-0.0008	0.9994
I(MCM6):I(DIAPH3.1)	-0.2374	0.7887	455.3889	-0.0005	0.9996
I(Contig20217.RC):I(DIAPH3.1)	-0.0848	0.9187	92.8253	-0.0009	0.9993
I(MTDH):I(Contig32125.RC)	-3.0599	0.0469	617.2640	-0.0050	0.9960

	Coef	HR	SE (coef)	z	Pr(>|z|)
I(MTDH):I(Contig32125.RC^3)	-40.4657	0.0000	3811.3305	-0.0106	0.9915
I(MCM6):I(Contig32125.RC)	0.2187	1.2444	1231.2267	0.0002	0.9999
I(Contig32125.RC):I(MCM6^2)	0.0539	1.0554	1627.7461	0.0000	1.0000
I(MCM6):I(Contig32125.RC^2)	-0.1021	0.9029	317.9007	-0.0003	0.9997
I(MCM6):I(Contig32125.RC^3)	1.3030	3.6804	12160.9196	0.0001	0.9999
I(Contig40831.RC):I(BBC3)	0.5622	1.7546	477.6828	0.0012	0.9991
I(BBC3):I(MS4A7)	0.0128	1.0129	237.0779	0.0001	1.0000
I(BBC3):I(PALM2.AKAP2)	0.0437	1.0447	53.5511	0.0008	0.9993
I(Contig20217.RC):I(BBC3)	0.2697	1.3095	211.3592	0.0013	0.9990
I(ORC6L):I(DIAPH3.2)	-0.1364	0.8725	191.1954	-0.0007	0.9994
I(MCM6):I(DIAPH3.2)	-0.4151	0.6603	796.6407	-0.0005	0.9996
I(MCM6^2):I(DIAPH3.2^2)	-0.0093	0.9907	2032.2013	0.0000	1.0000
I(Contig20217.RC):I(DIAPH3.2)	-0.0617	0.9401	429.4714	-0.0001	0.9999
I(CENPA):I(DIAPH3.2)	-0.0023	0.9977	94.6417	0.0000	1.0000
I(Contig40831.RC):I(RP5.860F19.3)	0.2566	1.2926	104.7481	0.0024	0.9980
I(RP5.860F19.3):I(EGLN1)	-0.3737	0.6882	95.8625	-0.0039	0.9969
I(C16orf61):I(GMPS)	-0.2814	0.7547	77.1363	-0.0036	0.9971
I(TGFB3):I(C16orf61)	1.1471	3.1490	0.0000	Inf	0.0000
I(C16orf61):I(MELK)	-0.0116	0.9884	0.0000	Inf	0.0000
I(DTL):I(C16orf61)	-0.2043	0.8152	0.0000	Inf	0.0000
I(MCM6):I(C16orf61)	-0.2696	0.7637	0.0000	Inf	0.0000
I(SCUBE2):I(AYTL2)	-0.3300	0.7190	0.0000	Inf	0.0000
I(RAB6B):I(SCUBE2)	0.0576	1.0593	0.0000	Inf	0.0000
I(TGFB3^2):I(SCUBE2)	-0.2261	0.7976	0.0000	Inf	0.0000
I(RUNDC1):I(FLT1)	-0.6345	0.5302	0.0000	Inf	0.0000
I(FLT1):I(Contig35251.RC)	0.5632	1.7562	0.0000	Inf	0.0000
I(WISP1):I(FLT1)	0.7370	2.0896	0.0000	Inf	0.0000
I(WISP1^3):I(FLT1)	8.3381	4180.3477	0.0000	Inf	0.0000
I(FLT1):I(AP2B1)	-1.5923	0.2035	0.0000	Inf	0.0000
I(FLT1):I(EGLN1)	-2.4188	0.0890	0.0000	Inf	0.0000
I(CDCA7):I(GNAZ)	-0.1695	0.8441	0.0000	Inf	0.0000
I(GNAZ):I(LGP2)	1.0317	2.8057	0.0000	Inf	0.0000
I(OXCT1):I(NMU)	0.8875	2.4290	0.0000	Inf	0.0000
I(CDC42BPA):I(MMP9)	0.0352	1.0358	0.0000	Inf	0.0000
I(RUNDC1):I(ECT2)	0.1662	1.1808	0.0000	Inf	0.0000
I(RUNDC1):I(LGP2)	0.6479	1.9115	0.0000	Inf	0.0000
I(MTDH):I(Contig35251.RC)	-0.0261	0.9742	0.0000	Inf	0.0000
I(TGFB3):I(Contig35251.RC)	0.2465	1.2795	0.0000	Inf	0.0000
I(ORC6L):I(Contig35251.RC)	-0.1256	0.8819	0.0000	Inf	0.0000
I(Contig35251.RC):I(ORC6L^2)	0.0572	1.0589	0.0000	Inf	0.0000
I(DCK):I(ECT2)	0.4943	1.6393	0.0000	Inf	0.0000
I(ECT2):I(GPR126)	-0.7516	0.4716	0.0000	Inf	0.0000
I(CDCA7):I(ECT2)	-0.2589	0.7719	0.0000	Inf	0.0000
I(MCM6):I(ECT2)	-0.4955	0.6092	0.0000	Inf	0.0000
I(EGLN1):I(GMPS)	0.8089	2.2454	0.0000	Inf	0.0000
I(RAB6B):I(KNTC2)	-0.2565	0.7738	0.0000	Inf	0.0000

	Coef	HR	SE (coef)	z	Pr(>|z|)
I(TGFB3):I(KNTC2)	0.0524	1.0538	0.0000	Inf	0.0000
I(TGFB3):I(KNTC2^2)	0.2041	1.2264	0.0000	Inf	0.0000
I(KNTC2^2):I(MELK^2)	-0.0129	0.9872	0.0000	Inf	0.0000
I(KNTC2):I(LOC643008)	-0.1428	0.8669	0.0000	Inf	0.0000
I(KNTC2):I(LOC643008^3)	-0.3483	0.7059	0.0000	Inf	0.0000
I(KNTC2):I(Contig20217.RC)	-0.1018	0.9032	0.0000	Inf	0.0000
I(WISP1):I(UCHL5)	0.6361	1.8891	0.0000	Inf	0.0000
I(WISP1^3):I(UCHL5)	3.5627	35.2573	0.0000	Inf	0.0000
I(WISP1):I(ORC6L)	0.3452	1.4122	0.0000	Inf	0.0000
I(WISP1^3):I(ORC6L)	0.1655	1.1800	0.0000	Inf	0.0000
I(WISP1):I(CDCA7)	0.0056	1.0056	0.0000	Inf	0.0000
I(WISP1):I(MCM6)	0.4825	1.6201	0.0000	Inf	0.0000
I(WISP1^3):I(MCM6)	0.7862	2.1951	0.0000	Inf	0.0000
I(WISP1):I(PITRM1^3)	5.5825	265.7406	0.0000	Inf	0.0000
I(WISP1^3):I(PITRM1^3)	6.7883	887.3967	0.0000	Inf	0.0000
I(WISP1):I(PRC1)	0.2792	1.3221	0.0000	Inf	0.0000
I(WISP1):I(ESM1)	0.0968	1.1017	0.0000	Inf	0.0000
I(WISP1^3):I(ESM1)	1.3130	3.7173	0.0000	Inf	0.0000
I(CDC42BPA):I(LOC643008)	0.3761	1.4565	0.0000	Inf	0.0000
I(CDC42BPA):I(LOC643008^3)	0.2471	1.2803	0.0000	Inf	0.0000
I(CDC42BPA^2):I(LOC643008^2)	-1.3624	0.2560	0.0000	Inf	0.0000
I(CDC42BPA^2):I(LOC643008^3)	-0.2646	0.7675	0.0000	Inf	0.0000
I(MCM6):I(SERF1A)	-0.4079	0.6650	0.0000	Inf	0.0000
I(RTN4RL1):I(AYTL2)	-2.2591	0.1044	0.0000	Inf	0.0000
I(AYTL2):I(C9orf30)	1.8189	6.1652	0.0000	Inf	0.0000
I(AYTL2):I(NMU)	0.5267	1.6933	0.0000	Inf	0.0000
I(LGP2):I(GSTM3)	0.1841	1.2021	0.0000	Inf	0.0000
I(RAB6B):I(UCHL5)	-0.2923	0.7465	0.0000	Inf	0.0000
I(RAB6B):I(Contig40831.RC)	-0.0527	0.9487	0.0000	Inf	0.0000
I(RAB6B):I(ORC6L)	-0.0836	0.9198	0.0000	Inf	0.0000
I(RAB6B):I(RFC4)	-0.2576	0.7729	0.0000	Inf	0.0000
I(RAB6B):I(CDCA7)	-0.0115	0.9885	0.0000	Inf	0.0000
I(RAB6B):I(C9orf30)	-0.3970	0.6723	0.0000	Inf	0.0000
I(RAB6B):I(PRC1)	-0.1447	0.8653	0.0000	Inf	0.0000
I(RAB6B):I(Contig20217.RC)	-0.3086	0.7345	0.0000	Inf	0.0000
I(RAB6B):I(EGLN1)	-0.9377	0.3915	0.0000	Inf	0.0000
I(HRASLS):I(RTN4RL1)	-0.2098	0.8107	0.0000	Inf	0.0000
I(RTN4RL1):I(PALM2.AKAP2)	0.0985	1.1035	0.0000	Inf	0.0000
I(MS4A7):I(UCHL5)	0.1644	1.1787	0.0000	Inf	0.0000
I(PALM2.AKAP2):I(PECI)	0.3973	1.4878	0.0000	Inf	0.0000
I(CDCA7):I(MTDH)	-0.0328	0.9678	0.0000	Inf	0.0000
I(MTDH):I(PITRM1)	-0.6765	0.5084	0.0000	Inf	0.0000
I(MTDH):I(PITRM1^2)	-1.8709	0.1540	0.0000	Inf	0.0000
I(MTDH):I(PITRM1^3)	-8.3174	0.0002	0.0000	Inf	0.0000
I(TGFB3):I(Contig40831.RC)	0.6448	1.9055	0.0000	Inf	0.0000
I(Contig40831.RC):I(ORC6L^2)	0.6172	1.8537	0.0000	Inf	0.0000
I(CDCA7):I(Contig40831.RC)	-0.1062	0.8993	0.0000	Inf	0.0000

	Coef	HR	SE (coef)	z	Pr(>|z|)
I(MCM6):I(Contig40831.RC)	-0.2011	0.8179	0.0000	Inf	0.0000
I(Contig40831.RC):I(PITRM1)	-0.3871	0.6790	0.0000	Inf	0.0000
I(Contig40831.RC):I(PITRM1^2)	-2.0889	0.1238	0.0000	Inf	0.0000
I(Contig40831.RC):I(PITRM1^3)	-4.6271	0.0098	0.0000	Inf	0.0000
I(CENPA):I(Contig40831.RC)	-0.3008	0.7402	0.0000	Inf	0.0000
I(TGFB3):I(DCK)	-0.6716	0.5109	0.0000	Inf	0.0000
I(TGFB3^2):I(DCK)	2.2072	9.0904	0.0000	Inf	0.0000
I(TGFB3^2):I(ORC6L)	1.4917	4.4448	0.0000	Inf	0.0000
I(TGFB3^2):I(RFC4)	0.4337	1.5430	0.0000	Inf	0.0000
I(CDCA7):I(TGFB3^2)	0.4906	1.6334	0.0000	Inf	0.0000
I(TGFB3):I(AP2B1)	-0.0116	0.9885	0.0000	Inf	0.0000
I(TGFB3):I(PITRM1)	0.0462	1.0473	0.0000	Inf	0.0000
I(TGFB3):I(PITRM1^3)	2.8509	17.3039	0.0000	Inf	0.0000
I(TGFB3):I(Contig20217.RC)	0.6052	1.8316	0.0000	Inf	0.0000
I(CENPA):I(MELK)	-0.0085	0.9915	0.0000	Inf	0.0000
I(STK32B):I(FBXO31)	0.4542	1.5749	0.0000	Inf	0.0000
I(DCK):I(GPR126)	-0.0826	0.9208	0.0000	Inf	0.0000
I(NM.004702):I(DCK)	0.7736	2.1675	0.0000	Inf	0.0000
I(HRASLS):I(FBXO31)	-0.1206	0.8863	0.0000	Inf	0.0000
I(PITRM1):I(FBXO31)	-0.6946	0.4993	0.0000	Inf	0.0000
I(PITRM1^3):I(FBXO31)	-5.8673	0.0028	0.0000	Inf	0.0000
I(CDCA7):I(SLC2A3)	0.0518	1.0531	0.0000	Inf	0.0000
I(LGP2):I(SLC2A3)	-0.2400	0.7866	0.0000	Inf	0.0000
I(PALM2.AKAP2):I(PECI.1)	0.6482	1.9120	0.0000	Inf	0.0000
I(ORC6L):I(LOC643008)	-0.2327	0.7924	0.0000	Inf	0.0000
I(ORC6L):I(LOC643008^3)	-0.1266	0.8811	0.0000	Inf	0.0000
I(ORC6L):I(Contig20217.RC)	-0.1045	0.9008	0.0000	Inf	0.0000
I(CDCA7):I(EGLN1)	0.6631	1.9408	0.0000	Inf	0.0000
I(CDCA7):I(NM.004702)	-0.2183	0.8039	0.0000	Inf	0.0000
I(PITRM1):I(LOC643008)	-0.1972	0.8211	0.0000	Inf	0.0000
I(LOC643008):I(PITRM1^2)	-0.3558	0.7006	0.0000	Inf	0.0000
I(PITRM1^3):I(LOC643008)	-3.0251	0.0486	0.0000	Inf	0.0000
I(PITRM1^3):I(LOC643008^3)	-0.4247	0.6539	0.0000	Inf	0.0000
I(Contig20217.RC):I(LOC643008)	-0.2397	0.7869	0.0000	Inf	0.0000
I(LOC643008):I(C20orf46)	-0.0561	0.9454	0.0000	Inf	0.0000
I(MCM6):I(Contig20217.RC)	-0.7534	0.4708	0.0000	Inf	0.0000
I(Contig20217.RC):I(IGFBP5)	-0.0342	0.9663	0.0000	Inf	0.0000
I(HRASLS):I(NM.004702)	-0.0264	0.9740	0.0000	Inf	0.0000
I(Contig20217.RC):I(PITRM1)	-0.1803	0.8351	0.0000	Inf	0.0000
I(Contig20217.RC):I(PITRM1^3)	-7.5199	0.0005	0.0000	Inf	0.0000
I(Contig20217.RC):I(IGFBP5.1)	-0.2156	0.8061	0.0000	Inf	0.0000
I(LGP2):I(NMU)	-1.3123	0.2692	0.0000	Inf	0.0000
I(CENPA):I(NM.004702)	-0.0821	0.9212	0.0000	Inf	0.0000

APPENDIX 16. Coefficients of the Principal Components from Principal Component Cox Regression Model – NKI70 Data

#	Comps	Coef	HR	SE (Coef)	z	Pr (>|z|)
1	1	0.8206	2.272	0.1345	6.10	0.0000
2	3	-0.6571	0.518	0.1664	-3.95	0.0001
3	4	0.9894	2.690	0.2458	4.03	0.0001
4	6	-0.4600	0.631	0.1576	-2.92	0.0035
5	7	-1.0694	0.343	0.2215	-4.83	0.0000
6	8	-0.6303	0.532	0.1919	-3.28	0.0010
7	9	1.1769	3.244	0.2148	5.48	0.0000
8	11	2.1300	8.415	0.4090	5.21	0.0000
9	14	-0.9535	0.385	0.3260	-2.92	0.0034
10	15	0.9370	2.552	0.2576	3.64	0.0003
11	16	1.1558	3.177	0.3048	3.79	0.0001
12	18	-0.8841	0.413	0.2708	-3.26	0.0011
13	19	0.9637	2.621	0.2922	3.30	0.0010
14	20	-3.1295	0.044	0.5488	-5.70	0.0000
15	22	2.2921	9.896	0.4202	5.45	0.0000
16	23	-0.8391	0.432	0.3232	-2.60	0.0094
17	24	-1.9186	0.147	0.3843	-4.99	0.0000
18	25	-1.3847	0.250	0.2999	-4.62	0.0000
19	26	1.6680	5.301	0.4253	3.92	0.0001
20	30	1.0285	2.797	0.3176	3.24	0.0012
21	32	-1.0131	0.363	0.3671	-2.76	0.0058
22	33	1.1420	3.133	0.3565	3.20	0.0014
23	34	1.1093	3.032	0.4567	2.43	0.0151
24	35	0.6931	2.000	0.3917	1.77	0.0768
25	36	1.2936	3.646	0.3556	3.64	0.0003
26	37	-1.2220	0.295	0.3528	-3.46	0.0005
27	38	-2.0139	0.133	0.4866	-4.14	0.0000
28	39	0.6964	2.007	0.4140	1.68	0.0926
29	41	1.4897	4.436	0.4293	3.47	0.0005
30	43	-1.8973	0.150	0.5369	-3.53	0.0004
31	44	-1.7497	0.174	0.6741	-2.60	0.0094
32	45	3.5577	35.081	0.6471	5.50	0.0000
33	46	-4.0079	0.018	0.7671	-5.23	0.0000
34	48	1.0739	2.927	0.4666	2.30	0.0214
35	49	1.7520	5.766	0.6286	2.79	0.0053
36	50	-2.0974	0.123	0.7139	-2.94	0.0033
37	51	-1.3793	0.252	0.5619	-2.45	0.0141
38	52	-2.9088	0.055	0.7725	-3.77	0.0002
39	53	1.1048	3.019	0.7014	1.58	0.1152
40	54	2.0178	7.522	0.6534	3.09	0.0020
41	55	4.7294	113.232	0.9922	4.77	0.0000

#	Comps	Coef	HR	SE (Coef)	z	Pr (>|z|)
42	57	3.6193	37.312	0.7986	4.53	0.0000
43	58	5.1328	169.485	1.1537	4.45	0.0000
44	62	-1.4051	0.245	0.8396	-1.67	0.0942
45	65	-2.4714	0.084	0.8577	-2.88	0.0040
46	67	-1.3446	0.261	0.7998	-1.68	0.0927
47	68	2.5310	12.566	1.1126	2.27	0.0229
48	69	-3.3712	0.034	1.3217	-2.55	0.0108

APPENDIX 17. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Linear Model – NKI70 Data

Total Covariates = 68

Factors	Diam.GT2	N.GE4	ER.Pos	Grade.Well	Grade.Intermediate	Age	TSPYL5	Contig63649.RC
Coef	-82.05	10.61	-136.02	52.06	3.96	-1.66	-104.26	7.91
SE.Coef	-10.293	1.154	-14.198	5.854	0.484	-2.269	-7.977	0.455
Factors	BBC3	DIAPH3.2	RP5.860F19.3	C16orf61	SCUBE2	EXT1	FLT1	GNAZ
Coef	-21.53	-179.69	155.63	-94.9	-68.06	-22.83	106.26	61.19
SE.Coef	-0.8	-7	8.297	-4.182	-9.071	-0.735	4.288	3.203
Factors	OXCT1	MMP9	RUNDC1	Contig35251.RC	ECT2	GMPS	KNTC2	WISP1
Coef	37.15	40.08	226.36	-58.84	46.29	-2.18	-218.54	95.82
SE.Coef	1.548	2.599	11.073	-2.868	2.259	-0.108	-10.721	4.427
Factors	CDC42BPA	SERF1A	AYTL2	GSTM3	GPR180	RAB6B	ZNF533	RTN4RL1
Coef	22.32	-87.91	20.6	17.89	-6.53	-112.31	-8.77	-180.76
SE.Coef	0.892	-3.216	0.699	1.289	-0.319	-8.334	-0.99	-8.049
Factors	UCHL5	PECI	MTDH	Contig40831.RC	TGFB3	MELK	COL4A2	DTL
Coef	-112.85	250.56	139	-109.37	-188.56	-150.04	470.21	4.74
SE.Coef	-3.865	11.37	5.345	-5.515	-11.77	-8.31	17.459	0.22
Factors	STK32B	DCK	FBXO31	GPR126	SLC2A3	PECI.1	ORC6L	RFC4
Coef	-325.03	23.4	27.9	-115.95	-293.68	-195.99	133.71	56.49
SE.Coef	-10.45	0.935	1.082	-8.169	-11.89	-8.055	8.491	2.399
Factors	CDCA7	LOC643008	MS4A7	MCM6	AP2B1	C9orf30	IGFBP5	HRASLS
Coef	76.05	-65.31	0.11	-40.73	-220.8	279.34	104.31	-109.48
SE.Coef	7.28	-5.256	0.007	-1.849	-9.277	9.035	9.307	-6.233
Factors	PITRM1	IGFBP5.1	NMU	PALM2.AKAP2	LGP2	PRC1	Contig20217.RC	CENPA
Coef	-277.04	84.26	37.39	-225.74	78.97	349	-144.92	8.74
SE.Coef	-11.876	6.56	3.157	-10.588	3.448	22.27	-7.448	0.69
Factors	EGLN1	NM.004702	ESM1	C20orf46				
Coef	-230.99	259.49	169.99	-88.68				
SE.Coef	-8.896	17.26	11.163	-5.234				

APPENDIX 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Total Covariates = 735

Comps	Coef	SE (Coef)	Comps	Coef	SE (Coef)	Comps	Coef	SE (Coef)
KNTC2^2:GPR126	-97.7188	-0.5675	FGF18:PITRM1^3	0	0	Diam.GT2	3.9982	0.5016
KNTC2:PECI.1	46.8334	0.4326	FGF18^2:PITRM1	-33.7028	-0.1488	N.GE4	12.3927	1.3473
KNTC2^2:PECI.1	0	0	FGF18^2:PITRM1^2	-31.5661	-0.0383	ER.Pos	-18.7695	-1.9592
KNTC2:CDCA7	62.4721	1.5443	FGF18^2:PITRM1^3	-946.7039	-0.451	Grade.Well	-4.1398	-0.4655
KNTC2^2:CDCA7	-19.8168	-0.1701	FGF18^3:PITRM1	417.4953	0.8224	Grade.Intermediate	0	0
KNTC2:LOC643008	-38.3922	-0.563	FGF18^3:PITRM1^2	-432.885	-0.2424	TSPYL5	0	0
KNTC2:LOC643008^2	-41.5101	-0.3596	FGF18^3:PITRM1^3	14545.7287	2.8917	DIAPH3	-9.0384	-0.5247
KNTC2:LOC643008^3	66.3639	0.4258	FGF18:NMU	37.8359	1.1009	NUSAP1	7.8253	0.5081
KNTC2^2:LOC643008	100.4678	0.4624	FGF18^2:NMU	0	0	AA555029.RC	-3.1053	-0.1311
KNTC2^2:LOC643008^2	1964.7364	4.7883	FGF18^3:NMU	218.8096	1.5853	ALDH4A1	7.7806	0.3125
KNTC2^2:LOC643008^3	-426.519	-0.807	DIAPH3.1:DIAPH3.2	114.7103	1.4873	BBC3	10.887	0.4048
KNTC2:IGFBP5	-82.9786	-1.6191	DIAPH3.1:DIAPH3.2^2	35.8133	0.2008	C16orf61	1.8095	0.0797
KNTC2^2:IGFBP5	-103.1645	-0.7402	DIAPH3.1^2:DIAPH3.2	78.9386	0.576	SCUBE2	-5.1281	-0.6835
KNTC2:IGFBP5.1	-113.2143	-1.8224	DIAPH3.1^2:DIAPH3.2^2	2351.6408	6.2798	EXT1	13.3908	0.4309
KNTC2^2:IGFBP5.1	-196.6756	-1.1319	DIAPH3.1:RUNDC1	-58.6247	-0.6988	FLT1	0	0
KNTC2:Contig20217.RC	-46.8601	-0.6955	DIAPH3.1^2:RUNDC1	88.8477	0.5081	GNAZ	0	0
KNTC2^2:Contig20217.RC	-82.348	-0.6233	DIAPH3.1:ECT2	82.4969	0.9937	OXCT1	10.2325	0.4263
KNTC2:NM.004702	41.8468	0.6217	DIAPH3.1^2:ECT2	-97.119	-0.6035	MMP9	2.7516	0.1784
KNTC2^2:NM.004702	0	0	DIAPH3.1:GMPS	9.3311	0.1016	RUNDC1	2.2977	0.1124
WISP1:UCHL5	144.952	0.9099	DIAPH3.1^2:GMPS	-301.6515	-1.341	Contig35251.RC	-2.6528	-0.1293
WISP1^2:UCHL5	0	0	DIAPH3.1:KNTC2	-7.8776	-0.1092	ECT2	9.249	0.4513
WISP1^3:UCHL5	1599.5789	0.7767	DIAPH3.1:KNTC2^2	-241.9536	-1.4561	GMPS	1.5373	0.0761
WISP1:DTL	-88.7939	-0.9999	DIAPH3.1^2:KNTC2	41.0234	0.2958	SERF1A	-4.3603	-0.1595
WISP1:DTL^2	117.3333	0.5572	DIAPH3.1^2:KNTC2^2	-65.6865	-0.1678	AYTL2	0	0
WISP1^2:DTL	-143.8727	-0.5448	DIAPH3.1:GSTM3	13.072	0.2478	GSTM3	0	0
WISP1^2:DTL^2	240.2054	0.3741	DIAPH3.1^2:GSTM3	-175.5305	-1.4698	GPR180	0	0
WISP1^3:DTL	-1595.3462	-2.2273	DIAPH3.1:UCHL5	67.884	0.6652	RAB6B	-10.4318	-0.7741
WISP1^3:DTL^2	5617.7505	3.342	DIAPH3.1^2:UCHL5	418.0169	1.9598	ZNF533	-15.4152	-1.7395
WISP1:STK32B	-95.7995	-0.742	DIAPH3.1:Contig40831.RC	-75.7193	-0.859	RTN4RL1	-21.5582	-0.9599
WISP1^2:STK32B	0	0	DIAPH3.1^2:Contig40831.RC	391.0657	1.9251	UCHL5	-29.9097	-1.0244
WISP1^3:STK32B	2539.3868	2.9336	DIAPH3.1:TGFB3	31.441	0.5284	PECI	-11.5073	-0.5222
WISP1:ORC6L	1.1953	0.0156	DIAPH3.1:TGFB3^2	-22.3771	-0.2235	MTDH	0	0
WISP1:ORC6L^2	50.299	0.2654	DIAPH3.1^2:TGFB3	-20.5927	0.1854	Contig40831.RC	2.1344	0.1076
WISP1^2:ORC6L	18.1527	0.0796	DIAPH3.1^2:TGFB3^2	255.7316	1.3099	COL4A2	29.7588	1.1049

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
WISP1^2:ORC6L^2	494.7829	0.8274	DIAPH3.1:MELK	32.341	0.5654	STK32B	-17.6141	-0.5663
WISP1^3:ORC6L	-287.9887	-0.4854	DIAPH3.1:MELK^2	164.7417	1.5955	DCK	-3.3918	-0.1355
WISP1^3:ORC6L^2	3983.863	2.8439	DIAPH3.1^2:MELK	150.9033	1.4675	FBXO31	0	0
WISP1:CDCA7	17.4104	0.4398	DIAPH3.1^2:MELK^2	-483.0455	-2.4162	GPR126	-4.8349	-0.3406
WISP1^2:CDCA7	102.1919	0.7912	DIAPH3.1:DTL	-13.8076	-0.1963	SLC2A3	-6.7252	-0.2723
WISP1^3:CDCA7	-564.4947	-1.844	DIAPH3.1:DTL^2	-583.9423	-3.8684	PECI.1	-30.3976	-1.2492
WISP1:MCM6	93.163	0.8582	DIAPH3.1^2:DTL	-165.6185	-1.1438	RFC4	6.4161	0.2725
WISP1:MCM6^2	63.4668	0.1981	DIAPH3.1^2:DTL^2	-1463.4064	-4.6658	CDCA7	3.1557	0.3021
WISP1^2:MCM6	-115.3485	-0.3254	DIAPH3.1:GPR126	39.2137	0.6809	MS4A7	-19.2357	-1.2988
WISP1^2:MCM6^2	1108.6304	1.0189	DIAPH3.1^2:GPR126	-323.074	-2.153	AP2B1	-3.1199	-0.1311
WISP1^3:MCM6	445.8361	0.4375	DIAPH3.1:ORC6L	-179.7445	-2.9728	C9orf30	12.1123	0.3918
WISP1^3:MCM6^2	0	0	DIAPH3.1:ORC6L^2	28.16	0.2212	IGFBP5	2.5687	0.2292
WISP1:PITRM1	-13.2986	-0.0979	DIAPH3.1^2:ORC6L	45.1629	0.3582	HRASLS	-2.9324	-0.1669
WISP1:PITRM1^2	-139.459	-0.3225	DIAPH3.1^2:ORC6L^2	1006.4237	3.4117	IGFBP5.1	2.3414	0.1823
WISP1:PITRM1^3	614.1252	0.5992	DIAPH3.1:RFC4	-50.9337	-0.6392	NMU	4.2538	0.3592
WISP1^2:PITRM1	-71.9616	-0.1513	DIAPH3.1^2:RFC4	40.5032	0.2712	PALM2.AKAP2	-3.3571	-0.1575
WISP1^2:PITRM1^2	0	0	DIAPH3.1:CDCA7	47.0944	1.4107	LGP2	2.7024	0.118
WISP1^2:PITRM1^3	0	0	DIAPH3.1^2:CDCA7	-23.0064	-0.296	PRC1	1.5501	0.0989
WISP1^3:PITRM1	-1924.2793	-1.3222	DIAPH3.1:MCM6	-176.5941	-2.1949	Contig20217.RC	0	0
WISP1^3:PITRM1^2	-2997.1046	-0.4449	DIAPH3.1:MCM6^2	-206.0322	-0.9905	CENPA	10.4201	0.8226
WISP1^3:PITRM1^3	16304.6286	0.6903	DIAPH3.1^2:MCM6	-171.8694	-1.0744	EGLN1	-18.2853	-0.7042
WISP1:PRC1	9.7906	0.1543	DIAPH3.1^2:MCM6^2	-1956.5308	-4.0606	NM.004702	23.6671	1.5742
WISP1^2:PRC1	-56.8093	-0.3165	DIAPH3.1:PITRM1	-0.3042	-0.0033	ESM1	22.5135	1.4785
WISP1^3:PRC1	-677.9993	-1.488	DIAPH3.1:PITRM1^2	0	0	Age	-0.3899	-0.5321
WISP1:ESM1	83.4844	1.2068	DIAPH3.1:PITRM1^3	0	0	Age^2	-0.0042	-0.4825
WISP1^2:ESM1	0	0	DIAPH3.1^2:PITRM1	307.5057	1.684	Age^3	-0.0001	-0.5031
WISP1^3:ESM1	935.1844	1.3267	DIAPH3.1^2:PITRM1^2	269.242	0.5297	Contig63649.RC	10.5845	0.6088
CDC42BPA:LOC643008	73.0075	1.207	DIAPH3.1^2:PITRM1^3	0	0	Contig63649.RC^2	50.7521	1.0308
CDC42BPA:LOC643008^2	38.4793	0.4905	DIAPH3.1:NMU	10.7562	0.2523	QSCN6L1	9.4715	0.5059
CDC42BPA:LOC643008^3	371.9915	4.3997	DIAPH3.1^2:NMU	49.335	0.4557	QSCN6L1^2	0	0
CDC42BPA^2:LOC643008	0	0	DIAPH3.1:Contig20217.RC	16.0426	0.2092	FGF18	-0.4282	-0.0287
CDC42BPA^2:LOC643008^2	-1892.1277	-7.3851	DIAPH3.1^2:Contig20217.RC	60.0875	0.3487	FGF18^2	12.3202	0.2862
CDC42BPA^2:LOC643008^3	1250.7532	4.6626	DIAPH3.1:CENPA	0.066	0.0016	FGF18^3	-1.789	-0.0275
SERF1A:CDCA7	-8.9753	-0.1623	DIAPH3.1^2:CENPA	-28.0871	-0.3355	DIAPH3.1	5.5927	0.3171
SERF1A:MCM6	-87.5069	-0.6165	Contig32125.RC:MTDH	-304.9577	-1.6393	DIAPH3.1^2	69.1009	1.286

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Total Covariates = 735

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
SERF1A:MCM6^2	-104.8996	-0.2926	Contig32125.RC^2:MTDH	200.5028	0.2451	Contig32125.RC	4.0167	0.1606
SERF1A:NMU	-34.0023	-0.3393	Contig32125.RC^3:MTDH	-5293.3471	-1.8279	Contig32125.RC^2	-28.1979	-0.2518
AYTL2:RTN4RL1	-291.1125	-1.6774	Contig32125.RC:CDCA7	0	0	Contig32125.RC^3	-139.2422	-0.525
AYTL2:STK32B	17.5435	0.1004	Contig32125.RC^2:CDCA7	208.0069	0.9952	DIAPH3.2	5.5086	0.2146
AYTL2:MS4A7	-69.3694	-0.6634	Contig32125.RC^3:CDCA7	-181.6365	-0.2798	DIAPH3.2^2	129.3841	1.4372
AYTL2:C9orf30	159.6552	0.7036	Contig32125.RC:MCM6	-16.1312	-0.1115	RP5.860F19.3	-0.2199	-0.0117
AYTL2:NMU	226.7086	2.7045	Contig32125.RC:MCM6^2	355.7805	0.872	RP5.860F19.3^2	-1.1595	-0.0173
GSTM3:IGFBP5.1	0	0	Contig32125.RC^2:MCM6	772.335	1.6477	KNTC2	-4.1829	-0.2052
GSTM3:LGP2	85.4494	1.2361	Contig32125.RC^2:MCM6^2	397.1189	0.2918	KNTC2^2	-85.2319	-1.3222
GPR180:MELK	120.8391	1.6089	Contig32125.RC^3:MCM6	1776.1413	1.3152	WISP1	-10.2087	-0.4717
GPR180:MELK^2	-214.7268	-1.5029	Contig32125.RC^3:MCM6^2	10463.0246	2.671	WISP1^2	39.1787	0.4178
RAB6B:UCHL5	94.6252	0.8718	BBC3:KNTC2	0	0	WISP1^3	0	0
RAB6B:Contig40831.RC	-21.7678	-0.4194	BBC3:KNTC2^2	0	0	CDC42BPA	0	0
RAB6B:ORC6L	-16.4576	-0.3039	BBC3:AYTL2	-38.3031	-0.2045	CDC42BPA^2	-185.9925	-1.4396
RAB6B:ORC6L^2	-374.7416	-2.8011	BBC3:Contig40831.RC	80.7813	0.6358	TGFB3	-12.2067	-0.7619
RAB6B:RFC4	-157.5693	-2.5949	BBC3:MS4A7	51.6454	0.6045	TGFB3^2	8.4138	0.1935
RAB6B:CDCA7	-4.653	-0.1888	BBC3:IGFBP5	-8.5115	-0.1199	MELK	-5.4244	-0.3004
RAB6B:C9orf30	-77.2637	-0.7717	BBC3:IGFBP5.1	-7.6291	-0.092	MELK^2	-40.6186	-0.8602
RAB6B:PRC1	-13.9237	-0.282	BBC3:PALM2.AKAP2	61.327	0.4225	DTL	-6.796	-0.3154
RAB6B:Contig20217.RC	-18.6004	-0.3518	BBC3:Contig20217.RC	72.5902	0.5452	DTL^2	-136.9075	-1.9548
RAB6B:EGLN1	-67.4989	-0.7792	DIAPH3.2:C16orf61	-355.6668	-2.6484	ORC6L	2.3318	0.1481
RTN4RL1:HRASLS	-78.1401	-0.8596	DIAPH3.2^2:C16orf61	200.3127	0.4927	ORC6L^2	112.1417	2.0738
RTN4RL1:PALM2.AKAP2	137.0885	1.1889	DIAPH3.2:ECT2	130.0405	1.0104	LOC643008	1.3497	0.1086
UCHL5:Contig40831.RC	37.4493	0.2587	DIAPH3.2^2:ECT2	-202.4861	-0.5891	LOC643008^2	-38.2016	-1.5578
UCHL5:TGFB3	20.1762	0.2032	DIAPH3.2:GMPS	20.1161	0.1634	LOC643008^3	11.958	0.4909
UCHL5:TGFB3^2	132.2538	0.7396	DIAPH3.2^2:GMPS	645.2223	2.0859	MCM6	-3.4154	-0.1551
UCHL5:MS4A7	24.2322	0.242	DIAPH3.2:KNTC2	-91.0748	-0.8127	MCM6^2	-23.4781	-0.3285
PECI:PALM2.AKAP2	77.5081	0.7639	DIAPH3.2:KNTC2^2	-468.9518	-1.676	PITRM1	-31.4602	-1.3486
MTDH:CDCA7	-33.5339	-0.6601	DIAPH3.2^2:KNTC2	204.6847	0.6293	PITRM1^2	0	0
MTDH:PITRM1	-145.6261	-0.8373	DIAPH3.2^2:KNTC2^2	2300.7817	2.4682	PITRM1^3	0	0
MTDH:PITRM1^2	-396.7722	-0.6571	DIAPH3.2:UCHL5	91.044	0.5824	C20orf46	-4.817	-0.2843
MTDH:PITRM1^3	-980.4209	-0.6666	DIAPH3.2^2:UCHL5	34.1264	0.0732	C20orf46^2	-83.8943	-2.7584
MTDH:EGLN1	0	0	DIAPH3.2:Contig40831.RC	-53.4619	-0.4207	Age:GNAZ	0.1968	0.4403
Contig40831.RC:TGFB3	70.4484	0.9553	DIAPH3.2^2:Contig40831.RC	763.3815	1.7943	Age^2:GNAZ	0.0048	0.4757
Contig40831.RC:TGFB3^2	0	0	DIAPH3.2:TGFB3	-27.6818	-0.298	Age^3:GNAZ	0.0002	0.9052

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
Contig40831.RC:MELK	53.2243	0.7318	DIAPH3.2:TGFB3^2	-2.3785	-0.0145	Age:Contig40831.RC	0.0497	0.1121
Contig40831.RC:MELK^2	115.7175	1.012	DIAPH3.2^2:TGFB3	-139.5277	-0.5372	Age^2:Contig40831.RC	0.0011	0.115
Contig40831.RC:STK32B	64.8538	0.4308	DIAPH3.2^2:TGFB3^2	497.8128	0.9633	Age^3:Contig40831.RC	0	0.1163
Contig40831.RC:ORC6L	-64.7193	-0.8172	DIAPH3.2:MELK	-207.7899	-2.4512	Contig63649.RC:RUNDC1	-132.5401	-1.5651
Contig40831.RC:ORC6L^2	108.4912	0.6182	DIAPH3.2:MELK^2	82.5907	0.547	Contig63649.RC^2:RUNDC1	-150.5721	-0.8688
Contig40831.RC:CDCA7	-31.5	-0.8568	DIAPH3.2^2:MELK	-16.7925	-0.0812	Contig63649.RC:WISP1	102.9121	1.2862
Contig40831.RC:MCM6	-193.3835	-2.0322	DIAPH3.2^2:MELK^2	-103.2196	-0.2332	Contig63649.RC:WISP1^2	153.2039	0.6426
Contig40831.RC:MCM6^2	69.8632	0.3685	DIAPH3.2:DTL	132.-.6991	-1.3692	Contig63649.RC:WISP1^3	520.5461	0.8332
Contig40831.RC:PITRM1	0.1159	0.001	DIAPH3.2:DTL^2	-1377.4237	-6.5623	Contig63649.RC^2:WISP1	138.8848	0.9669
Contig40831.RC:PITRM1^2	-206.4337	-0.6139	DIAPH3.2^2:DTL	-765.5927	-3.2794	Contig63649.RC^2:WISP1^2	153.2935	0.401
Contig40831.RC:PITRM1^3	-533.6747	-0.6827	DIAPH3.2^2:DTL^2	-388.6666	-0.7156	Contig63649.RC^2:WISP1^3	1988.8234	2.1768
Contig40831.RC:CENPA	-83.351	-1.2649	DIAPH3.2:ORC6L	-299.7651	-3.1973	Contig63649.RC:CDC42BPA	64.3536	0.5437
Contig40831.RC:NM.004702	-25.653	-0.3909	DIAPH3.2:ORC6L^2	154.8429	0.7648	Contig63649.RC:CDC42BPA^2	0	0
TGFB3:MELK	-21.0189	-0.3341	DIAPH3.2^2:ORC6L	-142.6393	-0.5366	Contig63649.RC^2:CDC42BPA	235.7999	0.6503
TGFB3:MELK^2	126.479	1.0757	DIAPH3.2^2:ORC6L^2	-95.1615	-0.1387	Contig63649.RC^2:CDC42BPA^2	-2758.7688	-1.6535
TGFB3^2:MELK	-9.4826	-0.0763	DIAPH3.2:CDCA7	90.499	1.9083	DIAPH3:QSCN6L1	-33.7478	-0.4569
TGFB3^2:MELK^2	-156.8209	-0.6475	DIAPH3.2^2:CDCA7	28.6905	0.1992	DIAPH3:QSCN6L1^2	16.4166	0.0912
TGFB3:DCK	-138.073	-1.5462	DIAPH3.2:MCM6	-86.1452	-0.7021	DIAPH3:GMPS	2.3973	0.0279
TGFB3^2:DCK	301.7028	1.5634	DIAPH3.2:MCM6^2	-280.3167	-0.9665	DIAPH3:GSTM3	17.3923	0.3272
TGFB3:ORC6L	64.1471	0.8577	DIAPH3.2^2:MCM6	112.5	0.3252	DIAPH3:GPR180	-92.7839	-1.1891
TGFB3:ORC6L^2	6.1481	0.0329	DIAPH3.2^2:MCM6^2	-1007.0352	-0.9975	DIAPH3:UCHL5	152.2034	1.4531
TGFB3^2:ORC6L	387.0997	2.1326	DIAPH3.2:LGP2	-203.4061	-1.4563	DIAPH3:Contig40831.RC	-18.7603	-0.2259
TGFB3^2:ORC6L^2	1059.07	1.6021	DIAPH3.2^2:LGP2	-675.1182	-1.4689	DIAPH3:MELK	40.0173	0.7062
TGFB3:RFC4	-17.7446	-0.2431	DIAPH3.2:Contig20217.RC	-21.1291	-0.1961	DIAPH3:MELK^2	-118.8973	-1.2379
TGFB3^2:RFC4	161.7783	1.1861	DIAPH3.2^2:Contig20217.RC	108.5091	0.3483	DIAPH3:DTL	-67.2033	-0.9812
TGFB3:CDCA7	3.5892	0.108	DIAPH3.2:CENPA	-27.9918	-0.4321	DIAPH3:DTL^2	-389.9479	-2.5379
TGFB3^2:CDCA7	81.6541	1.1003	DIAPH3.2^2:CENPA	314.4997	1.8467	DIAPH3:ORC6L	-51.6162	-0.8977
TGFB3:MCM6	9.6613	0.1277	DIAPH3.2:NM.004702	-38.9631	-0.4613	DIAPH3:ORC6L^2	58.3328	0.4816
TGFB3:MCM6^2	47.9936	0.2709	DIAPH3.2^2:NM.004702	298.6948	1.3418	DIAPH3:CDCA7	41.5252	1.3664
TGFB3^2:MCM6	-26.8702	-0.1722	RP5.860F19.3:Contig40831.RC	35.8748	0.4165	DIAPH3:MCM6	-41.2468	-0.5412
TGFB3^2:MCM6^2	-51.845	-0.1275	RP5.860F19.3^2:Contig40831.RC	-94.0872	-0.3826	DIAPH3:MCM6^2	50.2604	0.2953
TGFB3:AP2B1	-45.8701	-0.5083	RP5.860F19.3:MELK	25.1469	0.2995	DIAPH3:HRASLS	27.1838	0.4387
TGFB3^2:AP2B1	-136.7512	-0.7299	RP5.860F19.3:MELK^2	96.6531	0.5069	DIAPH3:PITRM1	-45.426	-0.4738
TGFB3:PITRM1	54.9153	0.6749	RP5.860F19.3^2:MELK	-141.4732	-0.5667	DIAPH3:PITRM1^2	-206.6671	-0.7182
TGFB3:PITRM1^2	-142.3809	-0.5857	RP5.860F19.3^2:MELK^2	-110.0866	-0.1718	DIAPH3:PITRM1^3	0	0

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
TGFB3:PITRM1^3	1009.6326	1.6221	RP5.860F19.3:RFC4	-26.331	-0.276	DIAPH3:CENPA	23.7247	0.6097
TGFB3^2:PITRM1	11.3793	0.074	RP5.860F19.3^2:RFC4	185.0258	0.757	NUSAP1:DIAPH3.2	-19.8153	-0.2714
TGFB3^2:PITRM1^2	747.786	1.597	RP5.860F19.3:EGLN1	-131.1723	-1.0545	NUSAP1:DIAPH3.2^2	65.2864	0.3542
TGFB3^2:PITRM1^3	-1274.1589	-1.0877	RP5.860F19.3^2:EGLN1	-174.465	-0.429	NUSAP1:RAB6B	-42.769	-0.9258
TGFB3:IGFBP5.1	0	0	RP5.860F19.3:NM.004702	-30.348	-0.4628	NUSAP1:TGFB3	42.2755	0.6499
TGFB3^2:IGFBP5.1	0	0	RP5.860F19.3^2:NM.004702	154.8823	0.7978	NUSAP1:TGFB3^2	0	0
TGFB3:Contig20217.RC	92.3892	1.1684	C16orf61:GMPS	-55.4151	-0.6121	NUSAP1:CDCA7	30.6749	1.181
TGFB3^2:Contig20217.RC	-25.7151	-0.1553	C16orf61:TGFB3	100.1529	0.9955	NUSAP1:MCM6	-7.7941	-0.1199
TGFB3:NM.004702	33.7118	0.5171	C16orf61:TGFB3^2	68.224	0.2861	NUSAP1:MCM6^2	155.3892	1.1054
TGFB3^2:NM.004702	-15.2328	-0.1089	C16orf61:MELK	-143.2906	-1.6678	NUSAP1:HRASLS	-110.0697	-2.1999
MELK:RFC4	53.7496	0.892	C16orf61:MELK^2	-194.6974	-1.3998	NUSAP1:CENPA	7.902	0.2564
MELK^2:RFC4	-93.8682	-1.0782	C16orf61:DTL	-132.6942	-1.2527	AA555029.RC:FLT1	162.3618	1.0833
MELK:CDCA7	17.7098	0.5407	C16orf61:DTL^2	-682.132	-2.3877	AA555029.RC:RTN4RL1	114.6038	0.8358
MELK^2:CDCA7	-21.7244	-0.3922	C16orf61:CDCA7	-18.681	-0.4929	ALDH4A1:RTN4RL1	-355.599	-2.6157
MELK:MCM6	-41.7907	-0.6519	C16orf61:MCM6	-104.0756	-1.0633	ALDH4A1:DTL	-293.1958	-2.2671
MELK:MCM6^2	103.3649	0.8604	C16orf61:MCM6^2	-189.1532	-0.9573	ALDH4A1:DTL^2	-281.6936	-0.6994
MELK^2:MCM6	45.2622	0.4719	C16orf61:HRASLS	-34.5055	-0.4766	ALDH4A1:AP2B1	-277.4297	-1.9442
MELK^2:MCM6^2	1005.9731	5.6791	SCUBE2:Contig35251.RC	0	0	QSCN6L1:DIAPH3.1	-22.5973	-0.2993
MELK:PITRM1	-31.6987	-0.3323	SCUBE2:AYTL2	-61.5381	-1.4032	QSCN6L1:DIAPH3.1^2	-109.364	-0.74
MELK:PITRM1^2	-384.9023	-1.4103	SCUBE2:RAB6B	16.601	0.7449	QSCN6L1^2:DIAPH3.1	0	0
MELK:PITRM1^3	0	0	SCUBE2:PECI	-0.53	-0.0165	QSCN6L1^2:DIAPH3.1^2	0	0
MELK^2:PITRM1	-335.7045	-1.7657	SCUBE2:TGFB3	5.7871	0.2588	QSCN6L1:DIAPH3.2	28.3291	0.2694
MELK^2:PITRM1^2	1289.3587	2.3192	SCUBE2:TGFB3^2	-60.859	-1.4184	QSCN6L1:DIAPH3.2^2	298.0364	1.2259
MELK^2:PITRM1^3	2327.0459	1.7306	SCUBE2:STK32B	-11.9572	-0.2036	QSCN6L1^2:DIAPH3.2	0	0
MELK:CENPA	-14.7952	-0.3423	SCUBE2:PECI.1	13.404	0.3671	QSCN6L1^2:DIAPH3.2^2	0	0
MELK^2:CENPA	40.0132	0.5334	SCUBE2:LOC643008	-1.6263	-0.0926	QSCN6L1:C16orf61	-107.6752	-0.9844
COL4A2:HRASLS	18.9023	0.1748	SCUBE2:LOC643008^2	-31.6697	-1.4	QSCN6L1^2:C16orf61	38.276	0.1165
COL4A2:LGP2	-15.314	-0.0929	SCUBE2:LOC643008^3	-14.3184	-0.6281	QSCN6L1:ECT2	-60.0986	-0.7006
DTL:MCM6	-55.835	-0.6479	FLT1:MMP9	0	0	QSCN6L1^2:ECT2	0	0
DTL:MCM6^2	46.2317	0.2471	FLT1:RUNDC1	-67.9311	-0.7868	QSCN6L1:KNTC2	-33.6772	-0.3729
DTL^2:MCM6	332.9411	1.8561	FLT1:Contig35251.RC	119.3718	1.029	QSCN6L1:KNTC2^2	-259.4247	-1.1291
DTL^2:MCM6^2	1121.3609	2.5294	FLT1:WISP1	135.2353	1.1861	QSCN6L1^2:KNTC2	35.852	0.1743
DTL:NMU	10.7136	0.2525	FLT1:WISP1^2	0	0	QSCN6L1^2:KNTC2^2	-1577.6957	-3.0611
DTL^2:NMU	268.3578	2.24	FLT1:WISP1^3	1689.8134	1.5401	QSCN6L1:SERF1A	-88.6391	-0.8038
STK32B:FBXO31	163.7173	0.9213	FLT1:AP2B1	-188.2553	-1.1256	QSCN6L1^2:SERF1A	48.8085	0.1701

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
STK32B:PRC1	74.6048	0.7589	FLT1:EGLN1	-401.1736	-2.4487	QSCN6L1:MTDH	-78.665	-0.7443
DCK:GPR126	-73.7734	-1.164	GNAZ:CDCA7	-56.9348	-1.457	QSCN6L1^2:MTDH	-92.95	-0.3817
DCK:SLC2A3	-99.3917	-0.6267	GNAZ:MS4A7	1.8221	0.0302	QSCN6L1:Contig40831.RC	-108.6936	-1.1288
DCK:NM.004702	155.1212	1.7705	GNAZ:LGP2	234.4414	2.1495	QSCN6L1^2:Contig40831.RC	-66.3465	-0.2524
FBXO31:HRASLS	-120.1101	-1.6273	OXCT1:IGFBP5.1	19.8032	0.2969	QSCN6L1:TGFB3	-16.0338	-0.2612
FBXO31:PITRM1	-101.0922	-0.59	OXCT1:NMU	219.4474	2.6252	QSCN6L1:TGFB3^2	115.8489	1.0301
FBXO31:PITRM1^2	-207.4046	-0.3359	MMP9:CDC42BPA	42.5382	0.6872	QSCN6L1^2:TGFB3	-78.0356	-0.5484
FBXO31:PITRM1^3	-2261.913	-1.1374	MMP9:CDC42BPA^2	47.6893	0.1599	QSCN6L1^2:TGFB3^2	127.6788	0.4588
SLC2A3:CDCA7	18.3129	0.399	MMP9:COL4A2	5.6042	0.0853	QSCN6L1:MELK	-92.7035	-1.1347
SLC2A3:LGP2	-65.9495	-0.5423	MMP9:IGFBP5	1.4489	0.0576	QSCN6L1:MELK^2	-277.7946	-1.809
PECI.1:RFC4	73.5036	0.5826	MMP9:IGFBP5.1	1.506	0.0521	QSCN6L1^2:MELK	-155.4548	-0.891
PECI.1:PALM2.AKAP2	75.967	0.6121	RUNDC1:Contig35251.RC	9.4245	0.0988	QSCN6L1^2:MELK^2	-1095.6019	-2.8029
ORC6L:LOC643008	-47.2238	-0.9807	RUNDC1:ECT2	42.7566	0.6089	QSCN6L1:DTL	-113.0127	-1.232
ORC6L:LOC643008^2	9.6843	0.1203	RUNDC1:TGFB3	20.5835	0.282	QSCN6L1:DTL^2	-301.0477	-1.4186
ORC6L:LOC643008^3	-166.6516	-1.5951	RUNDC1:TGFB3^2	-49.0432	-0.3552	QSCN6L1^2:DTL	61.9852	0.3103
ORC6L^2:LOC643008	-37.4535	-0.314	RUNDC1:LGP2	111.7442	0.9788	QSCN6L1^2:DTL^2	-1271.5936	-2.6203
ORC6L^2:LOC643008^2	192.0602	0.7295	Contig35251.RC:ECT2	-63.2376	-0.7072	QSCN6L1:FBXO31	0	0
ORC6L^2:LOC643008^3	-5.1729	-0.0147	Contig35251.RC:KNTC2	31.8978	0.265	QSCN6L1^2:FBXO31	172.9024	0.5338
ORC6L:Contig20217.RC	-20.6864	-0.2641	Contig35251.RC:KNTC2^2	-183.3753	-0.4651	QSCN6L1:ORC6L	-101.4148	-1.3233
ORC6L^2:Contig20217.RC	70.5814	0.4343	Contig35251.RC:CDC42BPA	65.8162	0.4765	QSCN6L1:ORC6L^2	-199.7039	-1.1687
ORC6L:CENPA	20.2398	0.5129	Contig35251.RC:CDC42BPA^2	-191.6038	-0.3371	QSCN6L1^2:ORC6L	93.2663	0.4928
ORC6L^2:CENPA	-175.9429	-2.1656	Contig35251.RC:RAB6B	-7.7143	-0.1475	QSCN6L1^2:ORC6L^2	126.4398	0.2804
RFC4:CDCA7	34.8242	0.7845	Contig35251.RC:ZNF533	0	0	QSCN6L1:RFC4	-38.2173	-0.3689
RFC4:MCM6	15.0328	0.1936	Contig35251.RC:MTDH	-118.2086	-0.8673	QSCN6L1^2:RFC4	210.1497	0.8725
RFC4:MCM6^2	102.8303	0.7213	Contig35251.RC:Contig40831.RC	41.2755	0.5469	QSCN6L1:CDCA7	-44.1621	-1.0281
RFC4:HRASLS	-60.2745	-1.1461	Contig35251.RC:TGFB3	60.7698	0.8647	QSCN6L1^2:CDCA7	67.9308	0.6973
RFC4:NM.004702	51.8753	0.6216	Contig35251.RC:TGFB3^2	-12.7675	-0.0918	QSCN6L1:MCM6	-79.3157	-0.8481
CDCA7:MCM6	6.7995	0.1744	Contig35251.RC:ORC6L	-58.2782	-0.6703	QSCN6L1:MCM6^2	-74.6837	-0.3406
CDCA7:MCM6^2	21.1042	0.2587	Contig35251.RC:ORC6L^2	119.8218	0.5073	QSCN6L1^2:MCM6	-19.4749	-0.0935
CDCA7:Contig20217.RC	11.1694	0.296	Contig35251.RC:CENPA	-53.0239	-0.7527	QSCN6L1^2:MCM6^2	-470.7939	-0.8947
CDCA7:EGLN1	73.589	1.3773	Contig35251.RC:NM.004702	-23.0237	-0.2957	QSCN6L1:Contig20217.RC	-70.1729	-0.6596
CDCA7:NM.004702	-33.4819	-1.1905	ECT2:KNTC2	84.176	0.89	QSCN6L1^2:Contig20217.RC	0	0
LOC643008:PITRM1	-26.4581	-0.372	ECT2:KNTC2^2	-112.0687	-0.5009	QSCN6L1:CENPA	-32.2611	-0.4593
LOC643008:PITRM1^2	-121.5492	-0.6132	ECT2:UCHL5	-110.8594	-0.8677	QSCN6L1^2:CENPA	221.1197	1.2638
LOC643008:PITRM1^3	-296.8736	-0.6756	ECT2:Contig40831.RC	0	0	QSCN6L1:NM.004702	-87.7566	-1.2704

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
LOC643008^2:PITRM1	60.4784	0.488	ECT2:TGFB3	56.1609	0.7976	QSCN6L1^2:NM.004702	10.2236	0.0565
LOC643008^2:PITRM1^2	0	0	ECT2:TGFB3^2	-0.4937	-0.0039	FGF18:SCUBE2	-3.8677	-0.1697
LOC643008^2:PITRM1^3	0	0	ECT2:MELK	156.3246	1.6774	FGF18^2:SCUBE2	-1.9213	-0.0376
LOC643008^3:PITRM1	-140.8889	-0.8346	ECT2:MELK^2	-254.8148	-1.2964	FGF18^3:SCUBE2	-2.3145	-0.0239
LOC643008^3:PITRM1^2	-750.7394	-1.5252	ECT2:DTL	102.8495	1.0658	FGF18:WISP1	-58.2511	-0.7605
LOC643008^3:PITRM1^3	-1908.6897	-1.4693	ECT2:DTL^2	172.3811	0.7664	FGF18:WISP1^2	-92.7376	-0.412
LOC643008:Contig20217.RC	-48.9463	-0.8129	ECT2:DCK	129.5741	1.0553	FGF18:WISP1^3	-271.7534	-0.3974
LOC643008^2:Contig20217.RC	22.1797	0.281	ECT2:GPR126	-71.0658	-1.1817	FGF18^2:WISP1	-42.8368	-0.2792
LOC643008^3:Contig20217.RC	-277.3443	-3.3589	ECT2:CDCA7	-44.0078	-1.0521	FGF18^2:WISP1^2	400.8055	0.7003
LOC643008:C20orf46	-9.6687	-0.2987	ECT2:MCM6	-106.9231	-1.0879	FGF18^2:WISP1^3	0	0
LOC643008:C20orf46^2	13.1688	0.3619	ECT2:MCM6^2	-63.8617	-0.2547	FGF18^3:WISP1	-207.8862	-0.6373
LOC643008^2:C20orf46	-14.2443	-0.4237	ECT2:HRASLS	-48.696	-0.561	FGF18^3:WISP1^2	-242.1502	-0.2323
LOC643008^2:C20orf46^2	-316.5247	-8.6526	ECT2:CENPA	18.7661	0.2934	FGF18^3:WISP1^3	0	0
LOC643008^3:C20orf46	-72.4348	-2.2607	GMPS:KNTC2	65.2943	0.7625	FGF18:TGFB3	-5.8925	-0.1061
LOC643008^3:C20orf46^2	574.8993	16.9853	GMPS:KNTC2^2	-86.6859	-0.3641	FGF18:TGFB3^2	10.5999	0.1051
MS4A7:LGP2	28.6231	0.3161	GMPS:DTL	11.1882	0.1236	FGF18^2:TGFB3	-52.8815	-0.4441
MCM6:PRC1	22.3473	0.2976	GMPS:DTL^2	681.8781	3.2611	FGF18^2:TGFB3^2	23.0723	0.0881
MCM6^2:PRC1	45.4486	0.2416	GMPS:CDCA7	21.8101	0.5561	FGF18^3:TGFB3	-154.9378	-0.5713
MCM6:Contig20217.RC	-132.1511	-1.3146	GMPS:MCM6	-8.828	-0.1144	FGF18^3:TGFB3^2	0	0
MCM6^2:Contig20217.RC	-99.6876	-0.4071	GMPS:MCM6^2	75.0393	0.4864	FGF18:STK32B	10.7677	0.0934
MCM6:NM.004702	-98.8302	-1.3399	GMPS:PRC1	51.7646	0.6498	FGF18^2:STK32B	-245.4121	-0.8935
MCM6^2:NM.004702	196.4661	1.0605	GMPS:Contig20217.RC	1.7394	0.02	FGF18^3:STK32B	-236.0006	-0.4395
IGFBP5:Contig20217.RC	3.2264	0.0669	GMPS:EGLN1	151.4902	1.4073	FGF18:DCK	-101.5225	-1.0296
HRASLS:Contig20217.RC	-7.5557	-0.1322	KNTC2:RAB6B	-7.5195	-0.1289	FGF18^2:DCK	-30.0975	-0.1209
HRASLS:NM.004702	-44.2932	-0.7029	KNTC2^2:RAB6B	-215.4752	-1.5822	FGF18^3:DCK	-377.7869	-0.6752
PITRM1:Contig20217.RC	-86.8835	-0.677	KNTC2:Contig40831.RC	-28.3785	-0.309	FGF18:CDCA7	4.2433	0.1462
PITRM1^2:Contig20217.RC	-67.777	-0.1214	KNTC2^2:Contig40831.RC	74.9621	0.2906	FGF18^2:CDCA7	-3.6679	-0.0525
PITRM1^3:Contig20217.RC	-2623.6751	-1.188	KNTC2:TGFB3	7.3703	0.0878	FGF18^3:CDCA7	15.989	0.1226
IGFBP5.1:Contig20217.RC	-4.094	-0.0724	KNTC2:TGFB3^2	135.1538	0.753	FGF18:MCM6	6.6259	0.0765
NMU:PALM2.AKAP2	9.4393	0.1467	KNTC2^2:TGFB3	794.6964	3.8419	FGF18:MCM6^2	69.8634	0.2879
NMU:LGP2	-200.5947	-2.1511	KNTC2^2:TGFB3^2	-30.4777	-0.0581	FGF18^2:MCM6	-38.0294	-0.192
NMU:PRC1	106.8049	3.4577	KNTC2:MELK	61.0656	0.8802	FGF18^2:MCM6^2	-215.864	-0.3822
NMU:Contig20217.RC	-25.3325	-0.3788	KNTC2:MELK^2	-193.2924	-1.5074	FGF18^3:MCM6	77.0567	0.1785
LGP2:C20orf46	-12.2246	-0.1081	KNTC2^2:MELK	-199.349	-1.0984	FGF18^3:MCM6^2	291.803	0.2571

Appendix 18. Coefficients for All Covariates Inversed Transformed from the Coefficients of the PLS Components from the PLS Cox Polynomial Model – NKI70 Data

Total Covariates = 735

Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)	Covs	Coef	SE (Coef)
LGP2:C20orf46^2	-2.2495	-0.006	KNTC2^2:MELK^2	-1524.5871	-3.8635	FGF18:PITRM1	98.4594	1.0466
CENPA:NM.004702	-75.7969	-2.1197	KNTC2:GPR126	8.3411	0.1352	FGF18:PITRM1^2	0	0

References

[1] Keselman H. J., Huberty C. J., Lix L. M., Olejnik S., Cribbie R., Donahue B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses.Review of Educational Research, 68, 350 –386.

[2] Maxwell S. E., & Delaney H. D. (2004).Designing experiments and analyzing data: A model comparison perspective (2nd ed.) Mahwah, NJ: Erlbaum.

[3] Wilcox R. R. (2005).Introduction to robust estimation and hypothesis testing (2nd ed.). New York: Elsevier.

[4] Breiman L., Friedman J.H., Olshen R. A., and Stone, C. J. (1984). Classification and regression trees. Wadsworth.

[5] Breiman, L. (2001), Random Forests, Machine Learning. 45(1), 5-32.

[6] Peter Buehlmann and Torsten Hothorn (2007), Boosting algorithms: regularization, prediction and model fitting. Statistical Science, 22(4), 477–505.

[7] Duo Zhou, Dinesh P. Mital, Shankar Srinivasan, and Syed Haque. (2011) Risk factors for open angle glaucoma - analyses using logistic regression. IJMEI 3(3):203-222.

[8] Duo Zhou, Dinesh P. Mital, and Shankar Srinivasan. (2013) Breast cancer diagnosis: a statistical analysis-based approach. IJMEI 5(4):321-333.

[9] Turnbull, B. and Weiss (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data.Biometrics34 (1978): 367-375.

[10] D'Agostino, Ralph B. (1970). Transformation to normality of the null distribution of g1. Biometrika 57 (3): 679–681.

[11] Anderson, T. W.; Darling, D. A. (1952). "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". Annals of Mathematical Statistics 23: 193–212.

[12] Smirnov N. (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19: 279–281.

[13] Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika 52 (3–4): 591–611.

[14] Therneau T., Grambsch P., and Fleming T. (1990) Martingale based residuals for survival models, Biometrika, March 1990.

[15] Wald, Abraham (1939). Contributions to the theory of statistical estimation and testing hypotheses. Annals of Mathematical Statistics. 10 (4): 299–326.

[16] Jelle J. Goeman (2012) Penalized R package, version 0.9-42.

[17] Jelle J. Goeman (2010) L1 penalized estimation in the Cox proportional hazards model, Biometrical Journal, 52 (1) 70-84.

[18] Balakrishnan N., Rao C.R., (2004). Handbook of Statistics. Advances in Survival Analysis. HandbNorth Holland; Handbook of Statistics (Book 23). 1st edition

[19] Kaplan, E. L.; Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assn. 53 (282): 457–481

[20] Greenwood, M. (1926) The natural duration of cancer. Reports on Public Health and Medical Subjects 33, 1–26. Her Majesty's Stationery Office, London.

[21] Hosmer, David, and Stanley Lemeshow (1999) Applied survival analysis: regression modeling of time to event data. (John Wiley & Sons, New York.

[22] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, Technometrics, 14, 945-965.

[23] Mantel Nathan (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50 (3): 163–70.

[24] Schoenfeld D (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 68: 316–319.

[25] Berty H. P.; Shi, H.; Lyons-Weiler, J. (2010). Determining the statistical significance of survivorship prediction models. J Eval Clin Pract 16 (1): 155–165.

[26] Mantel N, Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies. Journal of the National Cancer Institute. 22:719-748.

[27] Cox DR. (1972) Regression models and life tables. Journal of the Royal Statistical Society. B34:187-220.

[28] Bernstein L., Anderson J. and Pike M. C. (1981). Estimation of the proportional hazard in two-treatment-group clinical trials. Biometrics Vol. 37, No. 3 (Sep., 1981), pp. 513-519

[29] Peto R, Peto J. (1972). Asymptotically efficient rank invariant procedures. Journal of the Royal Statistical Society. A135:185-207.

[30] Gehan E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singlycensored samples. Biometrika 52:203-13.

[31] Breslow N. E. (1974). Covariance analysis of censored survival data. Biometrics. 30:89-99.

[32] Tarone R.E., Ware J. (1977) On distribution-free tests for equality of survival distributions. Biometrika. 64:165-60.

[33] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each subject. Part II: Analysis and Examples. British Journal of Cancer. 35:1-39.

[34] Kalbfleisch J.D., Prentice R.L. (1980) Statistical analysis of failure time data. New York: Wiley.

[35] Stanley S. (2005) Nonparametric survival analysis: Cox-Mantel tests and permutation tests. Unpublished Paper.

[36] Dabrowska D. M. (1987) Non-parametric regression with censored survival time data. Scandinavian Journal of Statistics, 14(3):181-197.

[37] Gonzalez-Manteiga W. and Cadarso-Suarez C. (1994) Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. Journal of Nonparametric Statistics, 4(1):65-78.

[38] Peña, E.A., Strawderman, R. and Hollander, M. (2001). Nonparametric estimation with recurrent event data. J. Amer. Statist. Assoc 96, 1299-1315.

[39] Ishwaran H. and Kogalur U.B. (2014). Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.5.3.

[40] Ishwaran H. and Kogalur U.B. (2007). Random survival forests for R. R News. v7/2. 25-31.

[41] Marsland S. (2009) Machine learning: An algorithmic perpective. Chapman & Hall/CRC Press. Taylor & Francis Group; Machine Learning & Pattern Recognition Series. 119-131.

[42] Hosmer D. W., Lemeshow S., May S. (2008) Applied Survival Analysis, Regression Modeling of Time to Event Data. Wiley-Interscience Publication, New York.

[43] Harrell F. E., Lee K. L. and Mark D. B. (2012) Tutorial in biostatistics / Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, Vol. 15, 361-387.

[44] Jenkins S.P. (2005a). Survival Analysis. Unpublished Lecture Notes manuscript, Institute for Social and Economic Research, University of Essex.

[45] Johnson N. L., Kotz S. (1970) Distributions in statistics: Continuous univariate distributions. Volume 1. Houghton-Mifflin, Boston [398, 399].

[46] Pan W. (2001) Using frailties in the accelerated failure time model. Lifetime Data Analysis; 7:55-64.

[47] Allison D.P. (1995). Survival analysis using SAS. Cary, NC: SAS Publishing.

[48] Anderson et al. (1993). statistical models based on counting processes. New York: Springer-Verlag.

[49] Cantor B. (2003). A SAS survival analysis techniques for medical research. Cary, NC: SAS Publishing.

[50] Feinleib M. (1960). A method of analyzing log normal distributed survival data with incomplete follow-up. Journal of the American Statistical Association, Vol.55: 534-545.

[51] Horner, R.D. (1987) Age at onset of Alzheimer's disease: Clue to the relative importance of etilogic factors? American Journal of Epidemiology. Vol.126: 409-414.

[52] Klein J.P., Pelz C., Zhang M. (1999) Modeling random effects for censored data by a multivariate normal regression model. Biometrics; 55:497-506.

[53] Keiding N., Andersen P. K., Klein J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. Statistics in Medicine.16 (1–3): 215–224.

[54] Bennett S. (1983). Log-Logistic regression models for survival data. Jounal of the Boyal Statistical Society, Series C 32 (2):165-171.

[55] Kimber A. C., Zhu C. (1999) Diagnostic for a Weibull Frailty model. In Statistical Inference and Design of Experiments, Dixit U, Satam M (eds). Narosa Publishing House: New Delhi, India; 36-46.

[56] Hougaard P. (1999) Fundamentals of Survival Data. Biometrics; 55:13-22.

[57] Wei L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in Survival Analysis. Statistics in Medicine 11 (14-15):1872-1879.

[58] Cox D.R. (1997). Some remarks on the analysis of survival data. First Seattle Symposium in Biostatistics: Survival Analysis, Springer: New York.

[59] Lambert P., Collett D., Kimber A., Johnson R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. Statistics in Medicine 23 (20): 3177–3192.

[60] Wang H., Dai H., Fu B. (2013) Accelerated failure time models for censored survival Data under referral bias. Biostatistics. 14(2):313-26.

[61] Dang U.J., McNicholas P.D. (2013) Accelerated failure time models for competing risks in a cluster weighted modelling framework. arXiv:1312.0859v1 [Stat.ME]: 1312.0859 Vol 1.

[62] Cox DR (1972). Regression models and life-tables (with discussion). Journal of Royal Statistical Society, Series B 34:187–220.

[63] Breslow, N. E. (1975). Analysis of survival data under the proportional hazards Model. International Statistical Review / Revue Internationale de Statistique 43 (1): 45–57.

[64] Lin D.Y. and Wei L.J. The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association, 84:1074-1078, 1989.

[65] Gui J, Li H (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21:3001–3008.

[66] J. F. Lawless. (2003) Statistical models and methods for lifetime data. Wiley.

[67] Wei L. J., Lin D. Y. and Weissfeld L. (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association, 84, 1065-1073.

[68] Andersen P. K. and Gill R. D. (1982) Cox's regression model for counting processes: a large sample study. Annals of Statistics, 10, 1100-1120.

[69] Prentice R. L., Williams B. J. and Peterson A. V. (1981) On the regression analysis of multivariate failure time data. Biometrika, 68, 373-379.

[70] Mogensen U.B., Ishwaran H., Gerds T.A. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software: Vol 50, 11.

[71] Strobl, C.; Boulesteix, A.L.; Kneib, T; Augustin, T.; and Zeileis A. (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics, 9(307).

[72] Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro and Mark Van Der Laan (2006). Survival Ensembles. Biostatistics, 7(3), 355--373.

[73] Goeman J.J. (2010). L-1 penalized Estimation in the Cox Proportional Hazards Model. Biometrical Journal 52 (1) 70-84.

[74] Cox, D. R.; Oakes, D. (1984). Analysis of Survival Data. New York: Chapman & Hall. ISBN 041224490X.

[75] Zhao, Qiang; Sun, Jianguo. (2007) Cox Survival Analysis of Microarray Gene Expression Data Using Correlation Principal Component Regression. Stat App in Genetics and Molecular Biology. Volume 6, Issue 1, 1544-6115, May 2007.

[76] Therneau T.M., Grambsch P. M. (2000). Modelling for survival data: Extending the Cox model. Statistics for Biology and Health Series. New York. Springer.

[77] Abrahamwics M., MacKenzie T. and Esdaile J.M. (1996) Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephristis. Journal of the American Medical Association, 91: 1432-1439.

[78] Herndon J.E. and Harrell F.E. (1995) The restricted cubic spline as baseline hazard in the proportion hazards model with step function time-dependent covariables. Statistics in Medicine, 14:2119-2129.

[79] Hess K.R. (1994) Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Statistics in Medicine, 13:1045-1062.

[80] Vaida F, Xu R. (2000) Proportional hazard model with random effects. Statistics in Medicine, 19 (24), 3309-3324.

[81] Rondeau V, Filleul L., Joly P. (2006) Nested frailty models using maximum penalized likelihodd estimation. Statistics in Medicine, 25 (23), 4036-4052.

[82] Rondeau V, Michiel S, Liquet B. Pignon J.P. (2008) Investigating trial and treatment heterogeneity in an individual subject data meta-analysis of survival data by means of the penalized maximum likelihood approach. Statistics in Medicine, 27 (11), 1894-1910.

[83] Little R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc. 90, 1112-1121.

[84] Li Q. H. and Lagakos S. W. (1997). Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. Statistics in Medicine16, 925–940.

[85] Ghosh D., Andlin D. Y. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. Biometrics: 59, 877–885.

[86] Lancaster T., Andintrator O. (1998). Panel data with survival: hospitalization of HIV-positive subjects. Journal of the American Statistical Association: 93, 46–53.

[87] Sinha D., Andmaiti T. (2004). A Bayesian approach for the analysis of panel-count data with dependent termination. Biometrics: 60, 34–40.

[88] Huang X., Andwolfe R. A. (2002). A frailty model for informative censoring. Biometrics: 58, 510–520.

[89] Liu L., Wolfe R. A., Andhuang X. (2004). Shared frailty models for recurrent events and a terminal event. Biometrics: 60, 747–756.

[90] Rondeau V, Commenges D., Joly P. (2003) Maximum penalized likelihood estimation in a Gamma-frailty model. Lifetime Data Analysis, 9 (2), 139-153.

[91] Mauguen A., Collette S., Pignon, J. P. and Rondeau V. (2013). Concordance measures in shared frailty models: application to clustered data in cancer prognosis. Statistics in Medicine 32, 27, 4803-4820

[92] Marquardt D. (1963). An algorithm for least-squares estimation of nonlinear parameters. SIAM. Journal of Applied Mathematics: 431-441.

[93] Wold, S.; Sjöström, M.; Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58 (2): 109–130

[94] Haenlein M., Kaplan, A. M. (2004). A Beginner's Guide to Partial Least Squares Analysis. Understanding Statistics **3** (4): 283–297

[95] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 2 (11): 559–572.

[96] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417-441, and 498-520.

[97] Hirschfeld, H.O. (1935) A connection between correlation and contingency. Proc. Cambridge Philosophical Society, 31, 520–524

[98] Marquardt, D. W. (1970) Generalized inverses, ridge regression, biased linear estimation and non-linear estimation, Technometrics, 12, 591–612.

[99] Tikhonov, A. N. (1943). On the stability of inverse problems. Doklady Akademii Nauk SSSR 39 (5): 195–198.

[100] Phillips, D. L. (1962). A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. Journal of the ACM 9: 84.

[101] Breiman, L. (1996) Heuristics of instability and stabilization in model selection.Ann. Statist.,24, 2350–2383.

[102] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58 (1): 267–288.

[103] Fu, W. (1998) Penalized regression: the bridge versus the lasso.J. Computnl Graph. Statist.,7, 397–416.

[104] Zou H., Hastie T. (2005) Regularization and variable selection via the elastic net. J.R. Statist. Soc. B 67, Part 2, PP. 201-320.

[105] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression.Ann. Statist.,32, 407–499.

[106] Wold S., Sjöström M., Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58 (2): 109–130.

[107] Trygg J., Wold S. (2002). Orthogonal Projections to Latent Structures. Journal of Chemometrics 16 (3): 119–128.

[108] Rokach L., Maimon O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc.

[109] Breiman L. (1984). *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC.

[110] Breiman L. (1996). Bagging predictors. Machine Learning 24 (2): 123–140.

[111] Breiman L. (2001). Random Forests. Machine Learning 45 (1): 5–32.

[112] Song L., Langfelder P., Horvath S. (2013) Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC Bioinformatics, 14:5

[113] Mogensen U.B., Ishwaran H., Gerds T.A. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software: Vol 50, 11.

[114] Therneau T., Atkinson B, Ripley B. (2014). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-5. http://CRAN.R-project.org/package=rpart

[115] Kuhfeld W. F. (1990) The PRINQUAL Procedure. SAS/STAT User's Guide, Fourth Edition, Volume 2, pp. 1265–1323.

[116] Van Houwelingen JC, Le Cessie S (1990) Predictive value of statistical models. Statistics in Medicine 8:1303–1325.

[117] Copas JB (1983) Regression, prediction and shrinkage. JRSS B 45:311–354.

[118] He X, Shen L (1997) Linear regression after spline transformation. Biometrika 84:474–481.

[119] Little RJA, Rubin DB (1987) Statistical analysis with missing data. New York: Wiley.

[120] Faris PD, Ghali WA, et al. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. J Clin Epidem 55:184–191, 2002.

[121] Hollander M. and Wolfe D.A. (1973). Nonparametric statistical methods. New York: Wiley.

[122] Press WH, Flannery BP, Teukolsky SA, Vetterling, WT (1988): Numerical recipes in C. Cambridge: Cambridge University Press.

[123] Sarle W. S. The VARCLUS Procedure. SAS/STAT User's Guide, 4th Edition, 1990. Cary NC: SAS Institute, Inc.

[124] Hoeffding W. (1948): A nonparametric test of independence. Ann Math Stat 19:546–57.

[125] Mardia, K. V., J. T. Kent and J. M. Bibby (1979). Multivariate Analysis. London: Academic Press.

[126] Venables, W. N. and B. D. Ripley (2002). Modern applied statistics with S, Springer-Verlag.

[127] David W. Roberts (2013). labdsv: Ordination and multivariate analysis for ecology. R package version 1.6-1. http://CRAN.R-project.org/package=labdsv

[128] Young F. W., Takane Y., Leeuw J. De. (1978). The principal components of mixed measurement level multivariate data: An alternativing least squares method with optimal scaling features. Psychometrika, 43:279-281.

[129] Harrell Jr F.E. (2014). rms: Regression Modeling Strategies. R package version 4.2-0. http://CRAN.R-project.org/package=rms

[130] Gareth Ambler and modified by Axel Benner (2010). mfp: Multivariable Fractional Polynomials. R package version 1.4.9.

http://CRAN.R-project.org/package=mfp

[131] Therneau T. (2014). A Package for Survival Analysis in S. R package version 2.37-7, <URL: http://CRAN.R-project.org/package=survival>.

[132] Therneau T. and Grambsch P. (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3.

[133] Hothorn T., Buhlmann P., Dudoit S., Molinaro A. and et al. (2006a). Survival Ensembles. Biostatistics, 7(3), 355–373.

[134] Hothorn T., Hornik K. and Zeileis A. (2006b). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651–674.

[135] Ishwaran H. and Kogalur U.B. (2013). Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.4.

[136] Ishwaran H. and Kogalur U.B. (2007). Random survival forests for R. R News 7(2), 25--31.

[137] Ishwaran H., Kogalur U.B., Blackstone E.H. and Lauer M.S. (2008). Random survival forests. Ann. Appl. Statist. 2(3), 841--860.

[138] Hothorn T., Buehlmann P., Dudoit S., Molinaro A. and Van Der Laan M. (2006). Survival Ensembles. Biostatistics, 7(3), 355--373.

[139] Strobl C., Boulesteix A.L., Zeileis A. and Hothorn T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25). URL <http://www.biomedcentral.com/1471-2105/8/25>.

[140] Strobl C., Boulesteix A.L., Kneib T., Augustin T. and Zeileis A. (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics, 9(307). URL http://www.biomedcentral.com/1471-2105/9/307

[141] Zou and Hastie (2005) Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B,76,301-320.

[142] Friedman J., Hastie T., Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/

[143] M.J. van de Vijver, Y.D. He, L.J. van 't Veer, H. Dai, A.A.M. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, and R. Bernards (2002). A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347 (25), 1999–2009.

[144] Hastie T., Robert T., Friedman J. (2009). The elements of statistical learning: Data mining Inference and Prediction. New York: Springer. pp. 485-586.

[145] Kolmogorov A (1933). Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari 4: 83–91.

[146] Smirnov N (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19: 279–281.

[147] Ambler G., and Royston P. (2001) Fractional polynomial model selection procedures: investigation of Type I error rate. Journal of Statistical Simulation and Computation 69, 89-108.

[148] Gerds & Schumacher (2006), Consistent estimation of the expected Brier score in general survival models with right-censored event times, Biometrical Journal 48, 1029-1040.

[149] SAS Institute Inc. 2011. Base SAS® 9.2 procedures Guide. Cary, NC: SAS Institute Inc.

[150] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

[151] Efron B., Tibshirani R. (1997) Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association, 92 (438), pp. 548-560. URL: http://links.jstor.org/sici?sici=0162-1459%28199706%2992%3A438%3C548%3AIOCT.B%3E2.0.CO%3B2-I

[152] Wehrens, Ron (2011) Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences (Use R!). Springer-Verlag Berlin Heidelberg 2011.

[153] Beer, D. G. et al. (2002) Gene-expression profiles predict survival of subjects with lung adenocarcinoma. Nature Medicine, 8, 816-824.

[154] Simon N., Friedman J., Hastie T., Tibshirani R. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software, 39(5). Url: http://www.stanford.edu/~hastie/Papers/v39i05.pdf.