DEVELOPMENT OF PROGNOSTIC MODEL FOR BREAST CANCER IN SHANGHAI BREAST CANCER SURVIVAL STUDY (SBCSS)

by

Run Fan

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

August 2015

Nashville, Tennessee

Approved:

Fei Ye, Ph.D. (Thesis Advisor)

Tatsuki Koyama, Ph.D.

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to the Biostatistics Graduate Program at Vanderbilt University. Here I have been fortunate to meet my thesis advisor, Dr. Fei Ye. She gave me the freedom to explore on my own, and meanwhile the guidance to help me overcome difficulties. Dr. Ye provides me with an excellent atmosphere for research. I would like to thank Dr. Jeffrey Blume and all the faculty members in Biostatistics Graduate Program. They provide strong fundamental trainings on statistical theories and applications and prepare me for a successful career. I would like to thank Dr. Frank Harrell for his amazing class, Regression Modeling Strategies. Many theories and strategies from his class were applied in this thesis. I would like to express my appreciation to my thesis committee, Dr. Tatsuki Koyama, for his encouragement and insightful comments. He not only helps me with my thesis work, but also helps me to grow into a researcher.

I also thank my family who unconditionally supported and encouraged me throughout the time of my study.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF APPENDIX

INTRODUCTION

Breast cancer severely threats women's health worldwide. In US, breast cancer is diagnosed in one in eight women in their lifetime and breast cancer ranks the second cancer-caused death for women after lung cancer [1]. It is estimated that, 231,840 invasive breast cancer cases will be diagnosed for 2015, which accounts for 29% of all cancers in women [1]. Despite of an overall favorable prognosis, the risk of breast cancer recurrence persists and can be high after 10 years from initial diagnosis. It is well recognized that breast cancer is a complex group of diseases with considerable inter-individual variability in prognosis. Understanding related prognostic factors and being able to more accurately predict outcomes of breast cancer would not only benefit breast cancer survivors and health care providers, but also have a major impact on public health policy development.

1. An accurate and reliable prognostic model is essential for decision making in breast cancer treatment.

Accurate prediction of breast cancer outcome facilitates the development of personalized treatment, reduces unnecessary adjuvant treatments, and results in effective and more precise decision making for both clinicians and patients. A reliable prediction model will be helpful when the patients and physicians need to decide whether certain adjuvant treatment and/or increasing soy intake after the initial surgery will be beneficial. If survival outcomes are likely to be significantly improved, then the benefit from the treatment and/or diet/lifestyle modification may outweigh the potential harm. On the other hand, if there is very little improvement in survival outcomes from interventions, the patient may decide not to go through them to avoid the possible side effects and toxicity.

For a long time, physicians determined the patients' recurrence chance and mortality solely based on their knowledge and experience with pathological factors including tumor size, tumor grade and lymph node status [2]. Patients may suffer from many side effects and toxicity from adjuvant treatments while the likelihood of improving their prognosis remains low, let alone quality of life [3]. Prognostic models greatly facilitate patients' and physicians' decision-making in clinical practice. With better prognostic models, we can improve the accuracy of patient's prognosis, weigh the potential benefits and risks of adjuvant treatment, and tailor adjuvant treatment based on predictive factors, especially for patients with invasive, early-stage breast cancer.

2. Currently available prognostic tools and models in breast cancer.

Many prognostic models and tools are developed to estimate survival of individual breast cancer patient, but very few accomplished sufficient reliability and generalizability for clinical use. The three most well established prognostic indexes and models are the Nottingham Prognostic Index (NPI), Adjuvant! and Predict [4].

Nottingham Prognostic Index is a prognostic scoring system developed with the Nottingham dataset [5]. It integrates tumor size, tumor grade and lymph node status information and categorizes patients into three groups with distinctive survival probabilities (low risk, intermediate risk and high risk) [5]. NPI has been validated with another dataset from Nottingham [6]. In 2009, NPI was improved to provide accurate individual estimate for survival probability in addition to prognostic scores [7].

Adjuvant! Online (http://www.adjuvantonline.com) is a web-based prognosis tool based on data from US Surveillance, Epidemiology & End Results Registry (www.seer.cancer.gov) [3, 7-10]. Adjuvant! determines 10-year recurrence and survival estimates based on information of age, tumor grade, tumor size, lymph node status, and estrogen receptor status [8]. Meanwhile, Adjuvant! calculates the efficacy of adjuvant treatment for each individual patient. Adjuvant! was shown accurate in the Dutch population during years 1987 and 1998 in estimation of overall survival (OS) and breast-cancer specific survival (BCSS) [7]. Furthermore, Adjuvant! was validated with data from British Columbia Cancer Agency (BCCS) and the United Kingdom [8, 11]. However, the prediction of efficacy of adjuvant systemic treatment was overestimated for younger patients, mainly because of relatively smaller sample size of younger population [12]. Moreover, the survival estimation by Adjuvant! is significantly biased downward for the treatment effect of endocrine therapy for pre-menopausal women who didn’t receive chemotherapy treatment [3].

PREDICT is an online web-based prognosis tool using the Eastern Cancer Registration and Information Centre (ECRIC) dataset in UK [13]. Models were built separately for estrogen receptor positive patients and negative patients, allowing us to observe the effect of estrogen receptor status overtime [14]. The prognostic factors included in the model are nodal status, tumor grade, tumor size, chemotherapy therapy, endocrine therapy and mode of tumor detection [13]. PREDICT was validated by an independent dataset from the West Midlands Cancer Intelligence Unit (WMCIU) of British Columbia, Canada [13]. In 2012, a new version of PREDICT, PREDICT Plus, was developed which incorporates human epidermal growth factor receptor 2 (Her2) status [15]. Validation was performed with the same Canadian dataset. PREDICT and PREDICT Plus provide better estimation for overall survival and breast cancer specific survival than Adjuavnt! [13, 15]. For patients with Her2 positive, Predict Plus performs the best among all three prognostic models [15].

3. Study objective

Most of validation on Adjuvant! and PREDICT were implemented with Western population. Given the huge difference of Asian and Western population in culture, genetic background and lifestyles, the models for breast cancer prognosis may vary dramatically. Adjuvant! has been shown to be overly optimistic about the survival of Asian breast cancer patients, especially for women under 40 years old [16]. A breast cancer cohort study for Asian women will fill this knowledge gap and provide us an ethnic specific evaluation for breast cancer prognosis. The objective of current study is to develop prognostic models to predict 5-year overall survival (5-year OS), 10-year overall survival (10-year OS), and 5-year breast cancer relapse-free survival (5-year RFS) for women treated for early breast cancer in an Asian population.

STUDY POPULATION AND VARIABLES

The study data are from Shanghai Breast Cancer Survival Study (SBCSS), which is a large, population-based cohort study [17]. This study includes 4,858 female participants, all of whom are the residents of Shanghai, China. Primary breast cancer cases were identified from Shanghai Cancer Registry. Patients were 20-75 years old at diagnosis and were recruited approximately 6 month after diagnosis. All patients were diagnosed between March 2002 and April 2006. There are three follow-up interviews at 18, 36, and 60 months after breast cancer diagnosis. For participants who were lost for follow-up interviews, their survival information was collected from Shanghai Vital Statistics Registry database. Survival data were collected over the time period 2003 – 2015 and is ongoing till the end of 2016. To this day, we have completed

5-year OS and 5-year RFS, as well as 10-year OS data. The exact survival time for some observations may be greater than five years (for 5-year overall and relapse-free survival models) or ten years (for 10-year overall survival model), which results in right-censoring.

The collection of variables included in prognostic model is critical for the model's performance. Single predictive factor rarely yields satisfactory prediction accuracy. Multiple risk factors are needed to reproducibly differentiate patients who response to treatment and those who do not. Adjuvant! Online prediction uses age, menopause status, tumor grade, tumor size, number of positive lymph node, and estrogen receptor (ER) status in the presence or absence of adjuvant treatments [3-4, 7, 9-12]. PREDICT incorporates information from age, nodal status, tumor grade, tumor size, ER status, chemotherapy therapy, endocrine therapy and mode of tumor detection [15]. PREDICT Plus incorporates Her2 information [15].

Prediction model from SBCSS dataset differs from other models: 1) it uses a well-established population-based Asian cohort, 2) it expands the selection of potential predictors to include progesterone receptor (PR) status, post-diagnostic weight change and modifiable lifestyle factors (physical activity and intake of soy protein). PR is a nuclear receptor that belongs to ovarian steroid hormone family that regulates target gene expression in response to progesterone, which is critical for the mammary gland development and function [18]. A Norwegian study showed that the survival probability of a high-risk patient with positive PR status is similar to an intermediate-risk patient [19]. Prediction model may gain more precision by including prognostic contribution of PR expression.

Post-diagnostic weight gain is prevalent among breast cancer patients as a side effect of systemic adjuvant treatment including chemotherapy and tamoxifen therapy [20]. Several studies have shown that post-diagnostic weight gain associates with poorer survival outcome, regardless of patients' age and menopausal status [21-28]. Change of energy intake, decrease in physical activity and lower level of metabolism are other contributing factors for post-diagnostic weight gain [25]. Post-diagnostic weight gain may cause obesity in some breast cancer patients. In a meta-analysis with ten-year follow-up, obese patients have 1.78 times more risk of disease recurrence and 1.36 times more risk of death [29]. Excess adiposity may have adverse effect on prognosis, possibly because it serves as a source for the generation of estrogen [30]. Indeed, estrogen concentration is substantially higher in obese women, which favors and mediates breast cancer growth [30]. Moreover, chemotherapy is known to be less efficient in obese patients as measured by the cell count of blood leukocyte [25].

Physical activity is important for maintaining good health in general as well as throughout breast cancer treatment. Inactivity has been shown to be a risk factor for the development of breast cancer. In addition, exercise capacity during cancer treatment may be reduced as a result of adverse side effects on the cardiopulmonary, neurologic, and muscular systems. Exercise can effectively reduce blood estrogen concentration by 15% to 25% [22]. Three MET-hours of physical activity benefits the prognosis of breast cancer patients, especially for patients who response to Tamoxifen treatment [23].

Moderate amount of soy proteins consumption is associated with improved breast cancer prognosis [31]. Isoflavones are a class of phytoestrogens which are plant derived compounds with estrogenic activity. Soy isoflavones may serve as an estrogen antagonist through binding of estrogen receptor and slow down cancer cell growth, especially for hormone-dependent cancers [32]. Soybeans and soy-derived food are the richest sources of isoflavones in human daily diet. According to study by Lu, et al., the hazard ratio of death for breast cancer patients who consume > 15.31 grams of soy protein per day is 0.71 [0.54-0.92] of those who consume ≤ 5.31 grams of soy protein per day [32]. The hazard ratio of breast cancer recurrence between these two groups of patients is 0.68 [0.54-0.87] [32]. However, higher doses (such as regularly consumption of high doses of soy protein powers and isoflavones) may increase the breast cancer progression.

Because of the importance of the above-mentioned factors, five groups of variables are considered in prognosis model of SBCSS dataset: (1) Demographic factors: age at diagnosis, post-diagnosis weight change, and BMI at 18-month interview; (2) Clinical factors: tumor grade (poorly differentiated, moderately differentiated, well differentiated), tumor size, number of positive nodes (>9, 4-9, 1-3, 0), and tumor-node-metastasis stage at diagnosis (III, IIB, IIA, I); (3) Pathological factors: ER status, PR status, and Her2 receptor status; (4) Lifestyle factors: physical activity at 18-month interview, soy protein intake at 18-month interview, and soy isoflavones at 18-month interview; (5) Primary and adjuvant treatment: mastectomy (radical mastectomy, other types of breast cancer surgery), chemotherapy (yes, no), radiotherapy (yes, no), and tamoxifen therapy (yes, no). Detailed information of these predictors' measurements in SBCSS dataset can be found in **Table 1-5**. Evaluation of predictive abilities of these potential variables will give us a more precise and less biased estimation of breast cancer patient survival outcomes.

MODELING STRATEGIES

1. Restricted Cubic Splines:

Among all 18 predictors, the following 7 are continuous variables: age at diagnosis, post-diagnosis weight change, BMI at 18-month interview, tumor size, exercise participation at 18-month interview, soy protein at 18-month interview, and soy isoflavones at 18-month interview. If a continuous predictor is presented in an ordinary linear regression model, the model assumes that the outcome behaves linearly in that predictor. However, this assumption may not always hold. Linearity may not be achieved even after predictor transformation. Spline function are polynomials for each interval of a predictor, which offers a precise function to the variable and keeps the estimation flexible in curve fitting [33]. In our study, restricted cubic splines are used to estimate the coefficients of continuous variables. Compared to smooth spline function, restricted cubic splines have less number of parameters (k-1) and behavior better in the tails. The restricted spline function with k knots t_1, …, t_k is stated in **equations 1-2** [33].

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{k-1} X_{k-1}$$ **[Equation 1]**

Where X1=X and for j=1, …, k-2,

$$X_{j+1} = (X - t_j)_+^3 - (X - t_{k-1})_+^3 (t_k - t_j)/(t_k - t_{k-1}) + (X - t_k)_+^3 (t_{k-1} - t_j)/(t_k - t_{k-1})$$ **[Equation 2]**

Due to sample size limitation, restricted cubic splines with 3 knots were used on the following three variables: age at diagnosis, post-diagnosis weight change, and isoflavones, selected based on scientific knowledge. The positions of knots were pre-specified as fixed equal-spaced quantiles: 10%, 50%, and 90% [33]. The three-knot restricted cubic spline function expands each continuous predictor into one non-linear term and two linear terms at both ends. The degree of freedom for each continuous predictor is 2. Indicator variables are used to expand categorical and binary predictors.

2. Proportion Hazard Assumption:

A multivariable survival model is used to estimate the prognosis treatment benefit for breast cancer patients and to compute the relationship between the time to event and a set of potential predictors. Cox proportional hazard model is a semi-parametric, multiplicative hazards model and is widely used for survival data analysis [34]. The basic form of Cox model is shown in **Equation 3** [34].

$$h(t|Z) = h_0(t)exp(\beta^t\mathbf{Z}) = h_0(t)exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + ... + \beta_4 Z_p)$$ **[Equation 3]**

$h_0(t)$ is an arbitrary baseline hazard rate. $\boldsymbol{\beta}$ is a vector of coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)^t$. **Z** is a vector of predictors. h(t|**Z**) is the hazard rate for any individual whose predictors equal to **Z** and time equals to t. The hazard ratio of two individuals, one with predictor vector Z and other with predictor vector Z*, is constant if both Z and Z* are fixed, regardless of change in time (**Equation 4**) [34].

$$\frac{h(t|\mathbf{Z})}{h(t|\mathbf{Z*})} = \frac{h_0(t)exp[\Sigma_{k=1}^{p}\beta_k Z_k]}{h_0(t)exp[\Sigma_{k=1}^{p}\beta_k Z_k^*]} = exp[\sum_{k=1}^{p}\beta_k(Z_k - Z_k^*)]$$ **[Equation 4]**

Proportional hazard assumption is a key assumption for Cox model and should be checked before using Cox regression model [33-34]. Schoenfeld residuals can be used to assess proportional hazard assumption [35]. Scaled Schoenfeld residuals for each predictor are calculated from Cox regression model based on their contribution to the partial log likelihood [35]. Plots were used to visualize the raw and spline-smoothed scaled Schoenfeld residuals over time. The slope of spline-smoothed Schoenfeld residuals for each predictor against time is zero under proportional hazard conditions. If the slope of spline-smoothed Schoenfeld residuals is non-zero against time for a variable, the proportional hazard assumption is deemed violated. In this case, we can either stratify the offending variable since the survival rate is different for at least one stratum or add time-dependent interaction for the variable.

3. Pre-specified Interactions:

In regression analysis, the effect of one predictor may differ for distinct levels of another predictor. Under this circumstance, an interaction term between the two predictors needs to be examined in the regression model. Six types of interactions that are common in clinical research are considered in this study [33]. (1) Interaction between different types of treatment (chemotherapy, radiotherapy, tamoxifen therapy, and mastectomy) and the severity of disease (tumor grade, tumor size, number of positive nodes, and tumor-node-metastasis stage at diagnosis). Treatment benefit may be more obvious for patients with severe breast cancer. (2) Interaction between age as diagnosis and life-style factors (level of exercise, intake of soy protein, intake of soy isoflavones). The lifestyle factors in our dataset are collected at the 18-month follow-up after diagnosis. The effect of life-styles factors is cumulative and chronic. The benefits of lifestyle factors may be more obvious for patients have longer exposure to these factors. (3) Interaction between age and different types of treatment (chemotherapy, radiotherapy, and tamoxifen therapy). Treatment effects are a balance between treatment benefits and side effects. The tolerance of side effects and toxicity varies significantly among patients from different age groups. (4) Interaction between estrogen and tamoxifen therapy. Tamoxifen is an antagonist of the estrogen receptor in mammary gland tissue. The benefit of tamoxifen therapy may be different for patients with different estrogen status. (5) Interaction between estrogen and soy protein intake (soyprotein and isoflavones). Soy foods are rich in isoflavones, which is a natural modulator of estrogen receptor. Because of the antiestrogenic property of isoflavones, it may provide beneficial effect for breast cancer patients. Similar as tamoxifen, the benefit of isoflavones may vary between estrogen receptor positive patients and estrogen receptor negative patients. (6) Interaction between age and severity of breast cancer. For patients with the same disease severity, their survival outcome might be very different if they belong to distinct age groups, and vice versa.

We tested above-mentioned interactions in all three survival models. Three interaction terms are significant by ANOVA test at the significance level of 5%: age*tumor-node-metastasis stage, age*soy isoflavones, and ER status*Tamoxifen therapy. These three interaction terms are included in all three models.

4. Collinearity and Redundancy Analysis:

Hierarchical cluster analysis on variables is performed to identify collinear predictors and redundancy. Spearman correlation is calculated with pairwise deletion of NAs (**Figure 1**). The graph shows that the correlation is about (1) 0.4 between variables ER status and PR status, (2) 0.6 between tumor-nodal-metastasis status and number of positive nodes, (3) 0.8 between BMI and post-diagnostic weight change, and (4) 0.9 between soy protein and soy isoflavones. If a predictor can be predicted from other predictor(s), coefficient estimates will be biased, standard errors of estimated coefficients will be inflated, and the power of the corresponding test will be reduced. Due to the high correlation between soy protein and soy isoflavones, variable soy_pro2 (Intake of soy protein) is removed from the full model without much loss of information. All treatment, clinical, and pathological predictors to include in the model are pre-specified based on scientific knowledge. The R^2 statistic of tumor-node-metastatic status by redundancy analysis did not change markedly after deletion of ER status and radiotherapy. These variables (tumor-node-metastatic status, ER status and radiotherapy) are generally considered important in breast cancer prognosis. Without causing concern of overfitting (see section 6), we included all these variables in the initial full model.

5. Missing Data Imputation:

There are three types of missing data, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR means that missing is completely unrelated to the responses. The missing information is just a random subset of the dataset. MAR means that missingness is not completely at random, rather it is related to the value of variables we measured. MCAR and MAR are called ignorable missing. MNAR means that the probability of missing is related to the response. MNAR is also called non-ignorable missing. It is reasonable to assume missingness in our dataset is missing at random (MAR), since most of them are due to loss to follow-up for reasons such as change of address and/or phone number, etc.

As aforementioned, data collection has been completed for 5-year and 10-year overall survival as well as relapse free survival. However, the disease relapse time information is missing for 63 patients (1.3%) who died of breast cancer. Compared with complete cases analysis, imputation yields less biased and more precise inferences. Replacing missing data with "best guess" values such as expected values from observed data result in biased estimates and underestimated variance. In 5-year relapse-free survival model, time-to-relapse for these patients were imputed with single imputation using "transcan" function in *Hmisc* R package. All 17 predictors, disease event indicator, relapse indicator, and survival time to relapse were used for the imputation. The imputed data was saved as the "complete" dataset for subsequent analyses.

In addition to outcome variables, missingness presents in predictors (**Table 6, Figure 2**). Her2 status, tumor grade and tumor size have large number of missing values (about 20%). ER status, PR status, nodal status, tumor-node-metastasis, exercise, soy protein, soy isoflavon, BMI and weight contain fewer than 10% missing data. Variables chemotherapy, tamoxifen therapy, radiotherapy, and age at diagnosis have no missing values. Cluster analysis shows that five pairs of variables, (1) Her2 status and tumor grade, (2) tumor size and number of positive nodes, (3) BMI and post-diagnostic weight change, (4) soy protein and soy isoflavones, (5) ER status and PR status, tend to be missing on the same patients (**Figure 3**). Multiple imputation (MI) was implemented using the "aregImpute" function in *Hmisc* R package for predictor imputation. MI imputes missing variables using the predictive mean matching method with weighted probability sampling of available data [36]. All 17 predictors were included in the imputation procedure, which was repeated for 10 times to account for variability introduced by imputation. All 10 imputed datasets were used in model fitting and validation. Coefficient estimates were averaged from 10 imputations. The variance-covariance matrix for parameter estimates was also adjusted for the variability introduced by MI. All imputed "complete" datasets were saved for later development of approximate models.

6. Overfitting:

When the prediction model is too complex with too many predictors, overfitting presents. The model describes random noise in the data rather than real signal from the underlying association. Heuristic shrinkage estimate is used to determine how reliable the model predicts new observations, whether overfitting is present, and whether data reduction is necessary. Heuristic shrinkage coefficient γ is the probability that the model predicts future data and is calculated from the total degrees of freedom of all predictors (p) and the model's chi-square statistic by likelihood ratio test (**Equation 5**) [33, 37].

$$\hat{\gamma} = \frac{model\chi^2 - p}{model\chi^2} \quad \textbf{[Equation 5]}$$

If heuristic shrinkage coefficient is greater than 0.9, the model has good calibration and no data reduction needed. For our models the heuristic shrinkage coefficient is 0.94 for 5-year OS model, 0.95 for 10-year OS model, and 0.95 for 5-year RFS model. When these models are validated on new datasets, only 5-6% of the model fitting is from noise.

We calculate the amount of complexity that our models can afford by calculating the effective sample size (m). A fitted survival model is reliable if the number of parameters is less than m/15 [33, 38-39]. In survival data, the effective sample size is the number of events (death or disease relapse). The effective sample sizes of 5-year OS, 10-year OS, and 5-year RFS model are 535, 950, and 845, respectively, among all 4,858 patients. As a rule of thumb, for each predictor included in a multivariable model at least 15 subjects (number of events for survival analysis) are required to avoid potential overfitting. The total number of degrees of freedom of all potential predictors is 36 (**Table 7**). No data reduction is needed.

7. Discrimination and Calibration:

Validation is critical for prognostic model development and for assessment of the model's reliability and generalizability. A prediction model needs to work not only for the dataset used to develop the model, it should also generalize to new datasets of the same target population. There are two types of validation: internal and external. Internal validation aims to correct for potential overfitting in the model. We internally validated our models using the .632 bootstrap method with 200 repetitions. The general performance and predictive accuracy of a prediction model are evaluated by discrimination and calibration.

Discrimination measures how well the prediction model separates the patients who experience the event from those who do not experience the event. Model discrimination in this study was measured using accuracy index D_{xy}, *c* statistic, and R^2. These statistics were calculated first using the original dataset and then corrected for overfitting. D_{xy} is used to measure the Somer's rank correlation between variables and binary outcome. If D_{xy} is close to 1, the model's discrimination ability is close to be perfect; if D_{xy} is close to 0, the model's estimation is random. Discrimination from internal validation is also measured by calculating area under receiver-operator-characteristic curve (ROC) (D_{xy}=2*(C-0.5)). Calibration describes the model's ability to measure the outcome and bias by comparing the estimated outcome with observed outcome. Calibration is measured by slope index (Slope). We also demonstrated the model's calibration ability graphically.

8. Approximation of the Full Model:

The full model is a multivariable Cox proportional hazard model that includes all 17 potential predictors and 3 interaction terms. A simplified model with high accuracy can be very convenient, especially when the full model is too complex for routine use. In case a simpler approximate model is desired, we performed model approximation using a fast backward step-down method. The full model with all 17 predictor and 3 interaction terms was against the fitted values from the full Cox regression model using ordinary least squares (OLS). The cutoff value for variables retained in the approximate model is α=0.1. To determine whether the approximate model has comparable accuracy as the full model, we fitted the approximate model against the full model's predicted values using OLS. The approximate model is considered to have comparable predictive accuracy as the full model if R^2 is greater than 0.95. The above procedure was repeated on all 10 imputed datasets. A variable was selected to be included in the final approximate model if it presented in more than 50% of the time.

The coefficients of the final approximate model are simply the average across all 10 fits.

The variance-covariance matrix of the final approximate model was calculated by **Equation 6** [33, 36]:

$$V = M^{-1}\sum_{i}^{M} Vi + \frac{M+1}{M}B$$ **[Equation 6]**

, where i denotes the i^{th} imputation (i=1,2,…,10), Vi is the variance-covariance matrix from the ordinary least square from the i^{th} imputed dataset, B is the between-sample variance of the variance-covariance matrix, and M is the total number of imputations. Using this method, we calculated the overall coefficient estimates, while adjusting for the variability introduced by multiple imputation for the final approximate model. We choose to use the imputed data from multiple imputation instead of single imputation to reduce the potential bias.

RESULTS

The outcome measurements in this study are 5-year OS, 10-year OS, and 5-year breast cancer RFS. Survival time was measured in years from the date of diagnosis to the last follow-up or event. Patients were considered as censored if the patient was lost to follow-up or if she was still alive five-years after diagnosis (for 5-year OS and 5-year RFS models) or ten-years after diagnosis (for 10-year OS model). Recurrence was defined as the reappearance of breast cancer either locally (the same place as original cancer), regional (near the surgery area) or distant (other area of body).

1. Results of Five-Year Overall Survival Model:

Slope of spline-smoothed Schoenfeld residuals is roughly zero against time for all variables in the full model (**Figure 4**). The effect of any predictor variable in our dataset is constant over time. PH assumption seems to hold, and Cox model is used to fit 5-year overall survival model. The likelihood ratio chi-squared value of full model is 579.23 with 36 degree of freedom. The shrinkage estimate is 0.94. So it is estimated that only about 6% of the model fitting is from noise. Coefficients of the full model are summarized in **Table 8**. The mathematical form of the full model is shown in the following equation (**Appendix Figure S1**). This equation can be used to precisely calculate the estimated log hazard ratio for a subject with a set of predictor values compare to the "reference" subject. Nomogram predicts median and mean survival time, based on the regression fit of full model (**Figure 5**). Nomogram can be used to manually calculate the predicted survival probabilities using this model. For each given value of predictor, the corresponding points can be read from the axis on the top of nomogram on a 0-100 scale for each predictor. Points from all predictors are added as “total points” to read the corresponding “median survival time” and “mean survival time”. Contribution of individual variable in predicting survival time is plotted by Wald chi-squared statistics, penalized for degree of freedom, in descending order (**Figure 6**). The ranking of importance of the predictors is nodal status, estrogen receptor status, tumor grade, age at diagnosis, post-diagnostic weight change, ER status*tamoxifen, Tamoxifen therapy, tumor-nodal-metastasis status, age*isoflavones, age*tumor-nodal-metastasis status, isoflavones, progesterone receptor status, tumor size, chemotherapy, radiotherapy, exercise, mastectomy, BMI, and Her2. Estimated hazard ratios for default setting of predictors is summarized (**Figure 7**). The dashed line indicated the survival time ratio equals to one. When age changes from its lower quartile to the upper quartile (46.4 year to 60.5 year), hazard ratio increases. The shaded bars represent 0.90, 0.95, 0.99 confidence limits. The strength of various predictors’ effect on log relative hazard was plotted (**Figures 8)**. All four treatments (chemotherapy, Tamoxifen, radiotherapy, and mastectomy) reduced the risk of death. Log relative hazard goes up for patients with more severe disease as measured by tumor size, tumor grade, number of positive nodes, and tumor-node-metastasis status. The effects of

Her2 and BMI are not obvious from the figure. For patients with tumor-node-metastasis stage I and IIA (tnm_status=2 and tnm_status=3), the log hazard rate first decreases then increases with age (**Figure 9**). However, for patients with tumor-node-metastasis stage IIB and III (tnm_status=4 and tnm_status=5), the log hazard rate goes up and then becomes flat with age. For ER positive patients, tamoxifen therapy reduces the relative hazard of death (**Figure 10**). To investigate the interaction of age and isoflavones, we categorized patients into four equal quantiles based on their age (<46.4, 46.4-51.1, 51.1-60.5, and >60.5). As the amount of isoflavones consumption increases, the log hazard ratio reduces for younger patients (**Figure 11**). However, for older patients (> 60.5 year), higher doses of isoflavone are associated with higher log hazard ratio, although the log hazard ratios of all age groups are below zero (**Figure 11**). The five-year overall survival rate is 89.0% (**Figure 12**).

Discrimination and calibration are quantified to assess the performance of our prognostic model (**Table 9**). The apparent Somers' D_{xy} is 0.533. The bias-corrected D_{xy} is 0.508. and the c-statistics is 0.754, suggesting that our model will likely discriminate well in future datasets. The original R^2 is 0.136 and the corrected R^2 is 0.126. The corrected slope shrinkage factor is 0.916, which is close to heuristic shrinkage estimate 0.94. We further determined the model's export-ability by optimism from Bootstrap, which is the average difference between the test quantity and training accuracy. The optimism for D_{xy}, R^2 and Slope are 0.025, 0.010, and 0.084, respectively. All of them are reasonably small, indicating that this model has a good export-ability.

The model's calibration accuracy in predicting 5-year OS was validated by bootstrap, using adaptive linear spline hazard regression. The black line is the observed mean events, the grey line indicates the ideal condition where slope equals to 1, and the blue line is generated by bootstrap datasets. The slope of observed versus predicted values is close to one, suggesting our model has good calibration (**Figure 13**).

Next, we developed an approximate model for 5-year OS data based on the fitted full model. Predictors BMI, mastectomy, and Her2 were removed from the full model since they presented in less than 50% of the time (**Figure 14**). We fitted each approximate model with different imputed dataset against the full model's predicted values and calculated R^2. All the R^2 are higher than 0.99, suggesting that the approximate model is quite accurate compared to the full model.

2. Results of Ten-Year Overall Survival

The slopes of spline-smoothed Schoenfeld residuals are constant around zero against time for all variables (**Figure 15**). PH assumption seems to hold and Cox model was used to fit 10-year OS model. The likelihood ratio chi-squared value of full model is 759.59 with 36 degree of freedom. The coefficients are summarized in **Table 11**. The shrinkage estimate is 0.95. So it is estimated that only 5% of the model fitting is from noise. The mathematical form of the simplified model is showed in **Appendix Figure S2**. Mean and median survival time can be calculated from nomogram (**Figure 16**). Based on chi-squared statistics, the ranking of importance of the predictors is number of positive nodes, age at diagnosis, tumor-nodal-metastasis status, tumor grade, age*tumor-nodal-metastasis status, post-diagnostic weight change, soy isoflavones, chemotherapy, age*isoflavones, tamoxifen therapy, tumor size, radiotherapy, ER status, exercise, ER status*tamoxifen, PR status, mastectomy, BMI, and Her2 (**Figure 17**). Estimated hazard ratios for default setting of predictors is summarized in **Figure 18**. The strength of various predictors’ effect on log relative hazard was plotted (**Figure 19-22**). The ten-year OS rate is 80.4% (**Figure 23**). The model's discrimination ability is measured by the original Somer's D_{xy}, the bias-corrected D_{xy}, and *c*-statistic. Their values are 0.466, 0.450, and 0.725, respectively (**Table 12**). The original R^2 is 0.151 and the corrected R^2 is 0.141. The corrected slope shrinkage factor is 0.937, which is close to heuristic shrinkage estimator 0.95. The optimism for D_{xy}, R^2, and slope are 0.016, 0.009, and 0.063, suggesting that the 10-year OS model has good export ability. In the calibration plot for 10-year OS model, the slope of observed versus predicted values is close to one, suggesting our model has good calibration (**Figure 24**).

Approximate model for 10-year overall survival model was developed. Predictors PR status, BMI, mastectomy, tumor size, and Her2 were removed from the full model (**Figure 25**). The approximate models from 10 imputed dataset explain at least 98% of the variance from the fitted value of full model.

3. Results of Five-Year Relapse-Free Survival

The slopes of spline-smoothed Schoenfeld residuals are constant around zero against time for all variables (**Figure 26**). PH assumption seems to hold and Cox model is used to fit 5-year RFS model. The likelihood ratio chi-square of full model is 653.53 with 36 degree of freedom. The coefficients are summarized in **Table 14**. The shrinkage estimate is 0.95. So it is estimated that only 5% of the model fitting is from noise. The mathematical form of the simplified model is showed in **Appendix Figure S3**. Mean and median survival time can be calculated from nomogram (**Figure 27**). Based on chi-square statistics, the ranking of importance of the predictors is number of positive nodes, age at diagnosis, tumor-nodal-metastasis status, tumor grade, ER status, age*tumor-nodal-metastasis status, post-diagnostic weight change, soy isoflavones, tamoxifen therapy, radiotherapy, ER status*tamoxifen therapy, tumor size, mastectomy, chemotherapy, age*isoflavones, exercise, PR status, BMI, and Her2 (**Figure 28**). Estimated hazard ratios for default setting of predictors is summarized in **Figure 29**. The strength of various predictors' effect on log relative hazard was plotted (**Figure 30-33**). The five-year relapse-free survival rate is 82.6% (**Figure 34**). The model's discrimination ability is measured by the original Somer's D_{xy} (0.453), the bias-corrected D_{xy} (0.434), and *c*-statistic (0.717) (**Table 15**). The corrected slope shrinkage factor is 0.93. The optimism for D_{xy}, R^2, and slope are 0.019, 0.011, and 0.074, suggesting that the 5-year RFS model has good export ability. In the calibration plot for 5-year RFS model, the slope of observed versus predicted values is close to one, suggesting our model has good calibration (**Figure 35**). Approximate model for 5-year RFS model was developed. Predictors PR status, BMI, and Her2 were removed from the full model (**Figure 36**). The approximate models from 10 imputed dataset explain more than 99% of the variance from the fitted value of full model.

SUMMARY

Breast cancer is the second most common cancer diagnosed among women worldwide [1]. Despite an overall favorable prognosis, considerable inter-individual variability in prognosis exists. Establishment of a prognosis prediction model would lay the foundation for the development of personalized treatment, which would maximize treatment efficacy, spare patients unnecessary treatment, reduce treatment-related toxicities, and identify women at high risk of recurrence for preventive intervention. There are widely used breast cancer prognosis tool based on European and North American populations, however, the prognosis tool for Asian population is understudied. We built a novel prognosis prediction model based on an Asian population-based cohort study. In addition to inclusion of the predictors in Adjuvant!, we expanded the prognostic model by incorporating lifestyle predictors (exercise level and intake of soy proteins), as well as PR status and Her2 status. Using Cox multivariate regression model, we established and internally validated prognostic models to predict 5-year OS, 10-year OS, and 5-year RFS. All three models are well-calibrated and showed good ability to discriminate among patients, with a *c*-statistic of 0.758, 0.728, and 0.721, respectively for 5-year OS, 10-year OS, and 5-year RFS. This predictive tool will be useful to facilitate breast cancer patients' decision-making, personalized treatment and recurrence prevention.

Complex statistical predictive models could be complicated for people from different fields. A nomogram elegantly reduces this complexity into simple numerical estimations of the probability of death or recurrence [40]. Nomogram may be tailored for an individual patients based on the their predictor information and greatly promotes our prognostic models' routine clinical use and increases its translational potential. Nomogam's graphical presentation is very practical and user-friendly for both physicians and patients. Nomogram also provides straightforward interpretation of risk assessment. The patient can visualize the benefit from certain treatment and/or change in their lifestyle and precisely estimate their prognostic. The input variables include patient demographic, clinical, pathological, treatment, and lifestyle information. Medium and mean survival/recurrence time were calculated.

The 5-year and 10-year OS approximate models were evaluated and compared to each other. Progesterone receptor status remained in 5-year OS approximate model, but not in the 10-year OS approximate model (**Figure 37**). The result about progesterone is intriguing because it might mean that the mechanisms of short-term and long-term disease outcomes are different and progesterone might be the key molecular that explains this difference. Progesterone is an ovarian steroid hormone. Progesterone and 17β-oestradiol are crucial in regulation of breast development and mammary carcinogenesis [41-43]. The function of 17β-oestradiol and progesterone are intimately linked together. Treatment by 17β-oestradiol alone failed to induce mammary gland cell proliferation in ovariectomized adult mice, but treatment by both 17β-oestradiol and progesterone induced sustained mammary gland cell proliferation [41, 44-45]. Mammary stem cells (MaSCs) are a very small subset of cells that can regenerate a functional mammary gland. Except developmental periods like puberty and pregnancy, MaSCs are at G0 resting phase and quiescent. So MaSCs are usually long-lived, which makes them vulnerable to accumulate mutagenesis. Misregulated expansion of MaSC pool may lead to abnormal cell growth and breast cancer initiation. Recent studies showed that MaSCs were important targets of steroid ovarian hormone signals and progesterone dramatically induced MaSC growth [46-47]. Progesterone may trigger breast cancer initiation through regulation of MaSCs growth, but may not be important for long-term disease progression.

Exercise level at 18-month after diagnosis was retained in both 5-year and 10-year OS approximate model (**Figure 37**). Very few studies assessed the association of post-diagnostic physical activity and mortality of breast cancer [48]. Physical activity impacts on breast cancer through three mechanisms: (1) physical activity lowers the concentration of circulating ovarian hormones, including estrogen and progesterone [49-51]; (2) physical activity reduces weight gain and obesity in breast cancer patients; (3) physical activity improves survival through reduction of insulin level [52]. Pharmacological antagonists of estrogen have been used in clinical practice since 1970s. Drugs interfere with progesterone signaling and the mitogenic effects of insulin are under development. Understanding the mechanism of physical activity in both short-term and long-term overall survival helps us to choose the best treatment strategy for breast cancer patients.

In summary, we built a novel breast cancer prognostic model among Asian women. Comparing to existing prognostic tools, we expanded the model by incorporating lifestyle predictors, PR and Her2 status. We performed internal validation on models developed for 5-year OS, 10-year OS, and 5-year RFS, and showed that all three models have satisfactory predictive accuracy. We further simplified the full models by performing model approximation using a fast backward step-down method. The approximate models will be convenient to practice use.

REFERENCES

1. American Cancer Society. *Cancer Facts & Figures*. 2015; Available from: http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2015/.
2. Cianfrocca, M. and L.J. Goldstein, *Prognostic and predictive factors in early-stage breast cancer.* Oncologist, 2004. **9**(6): p. 606-16.
3. Huober, J. and B. Thurlimann, *Adjuvant! When the new world meets the old world.* Lancet Oncol, 2009. **10**(11): p. 1028-9.
4. Wishart, G.C., et al., *A population-based validation of the prognostic model PREDICT for early breast cancer.* Eur J Surg Oncol, 2011. **37**(5): p. 411-7.
5. Haybittle, J.L., et al., *A prognostic index in primary breast cancer.* Br J Cancer, 1982. **45**(3): p. 361-6.
6. Todd, J.H., et al., *Confirmation of a prognostic index in primary breast cancer.* Br J Cancer, 1987. **56**(4): p. 489-92.
7. Mook, S., et al., *Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study.* Lancet Oncol, 2009. **10**(11): p. 1070-6.
8. Olivotto, I.A., et al., *Population-based validation of the prognostic model ADJUVANT! for early breast cancer.* J Clin Oncol, 2005. **23**(12): p. 2716-25.
9. Ravdin, P.M., et al., *Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer.* J Clin Oncol, 2001. **19**(4): p. 980-91.
10. Schmidt, M., et al., *Long-term outcome prediction by clinicopathological risk classification algorithms in node-negative breast cancer--comparison between Adjuvant!, St Gallen, and a novel risk algorithm used in the prospective randomized Node-Negative-Breast Cancer-3 (NNBC-3) trial.* Ann Oncol, 2009. **20**(2): p. 258-64.

11. Campbell, H.E., et al., *An investigation into the performance of the Adjuvant! Online prognostic programme in early breast cancer for a cohort of patients in the United Kingdom.* Br J Cancer, 2009. **101**(7): p. 1074-84.
12. Cufer, T., *Which tools can I use in daily clinical practice to improve tailoring of treatment for breast cancer? The 2007 St Gallen guidelines and/or Adjuvant! Online.* Ann Oncol, 2008. **19 Suppl 7**: p. vii41-5.
13. Wishart, G.C., et al., *PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer.* Breast Cancer Res, 2010. **12**(1): p. R1.
14. Azzato, E.M., et al., *Prevalent cases in observational studies of cancer survival: do they bias hazard ratio estimates?* Br J Cancer, 2009. **100**(11): p. 1806-11.
15. Wishart, G.C., et al., *PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2.* Br J Cancer, 2012. **107**(5): p. 800-7.
16. Bhoo-Pathy, N., et al., *Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients.* Eur J Cancer, 2012. **48**(7): p. 982-9.
17. Nechuta, S.J., et al., *The After Breast Cancer Pooling Project: rationale, methodology, and breast cancer survivor characteristics.* Cancer Causes Control, 2011. **22**(9): p. 1319-31.
18. Gao, X. and Z. Nawaz, *Progesterone receptors - animal models and cell signaling in breast cancer: Role of steroid receptor coactivators and corepressors of progesterone receptors in breast cancer.* Breast Cancer Res, 2002. **4**(5): p. 182-6.
19. Collett, K., R. Skjaerven, and B.O. Maehle, *The prognostic contribution of estrogen and progesterone receptor status to a modified version of the Nottingham Prognostic Index.* Breast Cancer Res Treat, 1998. **48**(1): p. 1-9.
20. Demark-Wahnefried, W., B.K. Rimer, and E.P. Winer, *Weight gain in women diagnosed with breast cancer.* J Am Diet Assoc, 1997. **97**(5): p. 519-26, 529; quiz 527-8.
21. Rock, C.L. and W. Demark-Wahnefried, *Nutrition and survival after the diagnosis of breast cancer: a review of the evidence.* J Clin Oncol, 2002. **20**(15): p. 3302-16.
22. Chlebowski, R.T., E. Aiello, and A. McTiernan, *Weight loss in breast cancer patient management.* J Clin Oncol, 2002. **20**(4): p. 1128-43.
23. Carmichael, A.R., *Obesity and prognosis of breast cancer.* Obes Rev, 2006. **7**(4): p. 333-40.
24. Protani, M., M. Coory, and J.H. Martin, *Effect of obesity on survival of women with breast cancer: systematic review and meta-analysis.* Breast Cancer Res Treat, 2010. **123**(3): p. 627-35.
25. Chen, X., et al., *Obesity and weight change in relation to breast cancer survival.* Breast Cancer Res Treat, 2010. **122**(3): p. 823-33.
26. Kroenke, C.H., et al., *Weight, weight gain, and survival after breast cancer diagnosis.* J Clin Oncol, 2005. **23**(7): p. 1370-8.
27. Caan, B.J., et al., *Post-diagnosis weight gain and breast cancer recurrence in women with early stage breast cancer.* Breast Cancer Res Treat, 2006. **99**(1): p. 47-57.
28. Caan, B.J., et al., *Pre-diagnosis body mass index, post-diagnosis weight change, and prognosis among women with early stage breast cancer.* Cancer Causes Control, 2008. **19**(10): p. 1319-28.
29. Goodwin, P., et al., *Multidisciplinary weight management in locoregional breast cancer: results of a phase II study.* Breast Cancer Res Treat, 1998. **48**(1): p. 53-64.

30. Clemons, M. and P. Goss, *Estrogen and the risk of breast cancer.* N Engl J Med, 2001. **344**(4): p. 276-85.
31. Rock, C.L., et al., *Nutrition and physical activity guidelines for cancer survivors.* CA Cancer J Clin, 2012. **62**(4): p. 243-74.
32. Shu, X.O., et al., *Soy food intake and breast cancer survival.* JAMA, 2009. **302**(22): p. 2437-43.
33. Harrell, F.E., *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2015: Springer.
34. Klein JP & Moeschberger ML, *Survival Analysis: Techniques for censored and truncated data* 2003: Springer.
35. Therneau, P.G.a.T., *Proportional hazards tests and diagnostics based on weighted residuals.* Biometrika, 1994. **81**: p. 515-26.
36. R. J. A. Little & D. B. Rubin, *Statistical Analysis with Missing Data*. second edition ed. 2002: Wiley, New York.
37. Van Houwelingen, J.C. and S. Le Cessie, *Predictive value of statistical models.* Stat Med, 1990. **9**(11): p. 1303-25.
38. Harrell, F.E., Jr., et al., *Regression modelling strategies for improved prognostic prediction.* Stat Med, 1984. **3**(2): p. 143-52.
39. Harrell, F.E., Jr., et al., *Regression models for prognostic prediction: advantages, problems, and suggested solutions.* Cancer Treat Rep, 1985. **69**(10): p. 1071-77.
40. Iasonos, A., et al., *How to build and interpret a nomogram for cancer prognosis.* J Clin Oncol, 2008. **26**(8): p. 1364-70.
41. Brisken, C., *Progesterone signalling in breast cancer: a neglected hormone coming into the limelight.* Nat Rev Cancer, 2013. **13**(6): p. 385-96.
42. Nandi, S., *Endocrine control of mammarygland development and function in the C3H/ He Crgl mouse.* J Natl Cancer Inst, 1958. **21**(6): p. 1039-63.
43. Lyons, W.R., *Hormonal synergism in mammary growth.* Proc R Soc Lond B Biol Sci, 1958. **149**(936): p. 303-25.
44. Beleut, M., et al., *Two distinct mechanisms underlie progesterone-induced proliferation in the mammary gland.* Proc Natl Acad Sci U S A, 2010. **107**(7): p. 2989-94.
45. Wang, S., L.J. Counterman, and S.Z. Haslam, *Progesterone action in normal mouse mammary gland.* Endocrinology, 1990. **127**(5): p. 2183-9.
46. Asselin-Labat, M.L., et al., *Control of mammary stem cell function by steroid hormone signalling.* Nature, 2010. **465**(7299): p. 798-802.
47. Joshi, P.A., et al., *Progesterone induces adult mammary stem cell expansion.* Nature, 2010. **465**(7299): p. 803-7.
48. Holmes, M.D., et al., *Physical activity and survival after breast cancer diagnosis.* JAMA, 2005. **293**(20): p. 2479-86.
49. Broocks, A., et al., *Cyclic ovarian function in recreational athletes.* J Appl Physiol (1985), 1990. **68**(5): p. 2083-6.
50. Bullen, B.A., et al., *Induction of menstrual disorders by strenuous exercise in untrained women.* N Engl J Med, 1985. **312**(21): p. 1349-53.
51. McTiernan, A., et al., *Adiposity and sex hormones in postmenopausal breast cancer survivors.* J Clin Oncol, 2003. **21**(10): p. 1961-6.

52. Goodwin, P.J., et al., *Fasting insulin and outcome in early-stage breast cancer: results of a prospective cohort study.* J Clin Oncol, 2002. **20**(1): p. 42-51.

Table 1: Descriptive Analysis of Demographic Predictors

```
age_diag : Age at diagnosis, years
       n  missing   unique    Info    Mean     .05     .10     .25     .50     .75     .90     .95
    4858        0      456       1   53.46    40.1    42.5    46.4    51.1    60.5    69.5    71.7

lowest : 20.4 22.6 23.3 23.9 24.0, highest: 74.6 74.7 74.8 74.9 75.0
```

```
bmi2 : BMI at 18-month interview, kg/m2
       n  missing   unique    Info    Mean      .05      .10      .25      .50      .75      .90      .95
    4405      453     1282       1   24.45    19.49    20.42    22.07    24.09    26.50    28.83    30.44

lowest : 13.74 14.57 14.95 15.07 15.43
highest: 39.91 39.93 40.51 43.82 48.51
```

```
Postweight_change : Post-diagnosis weight change between first and second interview, kg
       n  missing   unique    Info     Mean   .05   .10   .25   .50   .75   .90   .95
    4405      453       74    0.99   0.7883    -5    -3    -1     1     3     5     6

lowest : -24 -22 -21 -19 -17, highest:  14  15  16  22  53
```

Table 2: Descriptive Analysis of Pathological Predictors

```
erstatus : estrogen receptor status, 1=positive, 0=negative
       n  missing   unique    Info    Sum     Mean
    4795       63        2    0.69   3089   0.6442

prstatus : progesterone receptor status, 1=positive, 0=negative
       n  missing   unique    Info    Sum     Mean
    4781       77        2    0.73   2799   0.5854

her_2 : her_2 receptor status, 1=negative, 2=borderline, 3=positive
       n  missing   unique    Info     Mean
    3881      977        3    0.71   0.6393

0 (2492, 64%), 1 (297, 8%), 2 (1092, 28%)
```

Table 3: Descriptive Analysis of Clinical Predictors

```
hgrade : Tumor grade, 1=well differentiated, 2=moderately differentiated, 3=poorly differentiated
       n  missing   unique    Info    Mean
    3865      993        3    0.83   2.212

1 (579, 15%), 2 (1888, 49%), 3 (1398, 36%)
```

```
tumorsize : Tumor size, centimeters
       n  missing   unique    Info    Mean   .05   .10   .25   .50   .75   .90   .95
    4121      737       28    0.92   2.909     1     1     2     2     3     4     5

lowest :  1  2  3  4  5, highest: 26 27 28 29 30

nodal_status: Number of positive nodes, 1=0, 2=1-3, 3=4-9, 4=>9
       n  missing   unique    Info     Mean
    4577      281        4    0.72   0.5353

0 (2968, 65%), 1 (977, 21%), 2 (423, 9%), 3 (209, 5%)

tnm_status : Tumor-Node-Metastasis stage at diagnosis, 2=I, 3=IIA, 4=IIB, 5=III
       n  missing   unique
    4628      230        4

2 (1679, 36%), 3 (1646, 36%), 4 (837, 18%), 5 (466, 10%)
```

Table 4: Descriptive Analysis of Treatment Predictors

```
chemotherapy : Received Chemocherapy, 1=yes, 0=no
        n  missing   unique     Info      Sum     Mean
     4858        0        2     0.22     4478   0.9218
```

```
radiotherapy : Received Radiotherapy, 1=yes, 0=no
        n  missing   unique     Info      Sum     Mean
     4858        0        2     0.66     1587   0.3267
```

```
tamox : Received Tamoxifen therapy, 1=yes, 0=no
        n  missing   unique     Info      Sum     Mean
     4858        0        2     0.75     2511   0.5169
```

```
bc_surgery : Received Mastectomy, 1=radical mastectomy, 0=other types of BC surgery
        n  missing   unique     Info      Sum     Mean
     4849        9        2     0.18     4530   0.9342
```

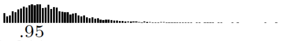
Table 5: Descriptive Analysis of Lifestyle Predictors

```
exercise2 : Exercise participation at 18-month interview, MET-hours per week
```

```
        n  missing   unique     Info     Mean      .05      .10      .25      .50      .75      .90      .95
     4413      445      878     0.98    13.37     0.00     0.00     0.00    10.00    19.80    32.40    42.57

lowest :   0.0000   0.3167   0.5000   0.6083   0.6333
highest: 105.0000 106.4000 109.2000 113.9417 141.2000
```

```
soy_pro2 : Soy protein at 18-month interview, grams per day
```



```
        n  missing   unique     Info     Mean      .05      .10      .25      .50      .75      .90      .95
     4407      451     4375        1    12.47    2.030    3.476    6.529   10.981   16.635   23.079   27.844

lowest :  0.00000  0.01232  0.01901  0.03467  0.04752
highest: 56.48474 59.11909 60.19235 67.11672 70.63788
```

```
isoflavon2 : Soy isoflavones at 18-month interview, mg per day
```

```
        n  missing   unique     Info     Mean      .05      .10      .25      .50      .75      .90      .95
     4407      451     4375        1    48.98    6.531   11.733   23.732   41.432   65.269   93.276  115.847

lowest :   0.00000   0.02963   0.07408   0.10017   0.11146
highest: 251.43137 258.38586 348.20136 355.67427 408.27069
```

Table 6: Summary of Missingness of All Predictors

Predictor	Variable Name	Number of Missing (Percent)
Demongraphic Variables:		
Age	agediag	0 (0)
BMI at 18-month	bmi2	453 (9.32%)
Weight Change	Postweight-change	453 (9.32%)
Pathological Variables:		
Tumor Grade	hgrade	993 (20.44%)
Tumor Size	tumorsize	737 (15.17%)
Number of Positive Nodes	nodalstatus	281 (5.78%)
Tumor-Node-Metastasis Stage	tnmstatus	230 (4.73%)
Clinical Variables:		
Estrogen Receptor Status	erstatus	63 (1.30%)
Progesterone Receptor Status	prstatus	77 (1.59%)
HER2 Status	her2	977 (20.11%)
Treatment Variables:		
Chemocherapy	chemotherapy	0 (0)
Radiotherapy	radiotherapy	0 (0)
Tamoxifen Therapy	tamox	0 (0)
Mastectomy	bcsurgery	9 (0.19%)
Lifestyle Variables:		
Exercise	exercise2	445 (9.16%)
Soy Protein	soypro2	451 (9.28%)
Soy Isoflavon	isoflavon2	451 (9.28%)

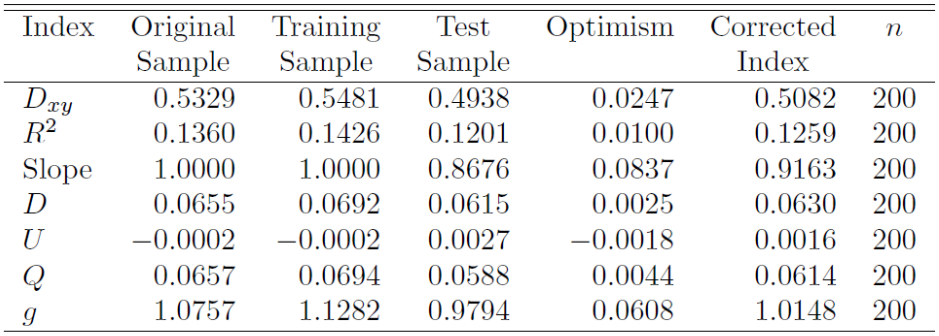
Table 7: Degree of Freedom for Candidate Predictors

	Type	Degree of freedom
Demographic Predictors:		
age at diagnosis	Continuous	2
BMI at 18-month interview	Continuous	1
Post-diagnostic weight change	Continuous	2
Pathological Predictors:		
Estrogen Receptor status	Binary	1
Progesterone Receptor status	Binary	1
HER2	Categorical	2
Clinical Predictors:		
Tumor grade	Categorical	2
Tumor Size	Continuous	1
Number of Positive Nodes	Categorical	3
Tumor-Node-Metastasis stage (TNM)	Categorical	3
Treatment Predictors:		
Chemotherapy	Binary	1
Radiotherapy	Binary	1
Tamoxifen	Binary	1
Mastectomy	Binary	1
Lifestyle Predictors:		
exercise at 18-month interview	Continuous	1
Soy isoflavones2 at 18-month interview	Continuous	2
Interaction terms:		
ER status * Tamoxifen	/	1
Age * TNM status	/	6
Age * Isoflavones	/	4
Total:		36

Table 8: Summary of 5-year Overall Survival Model

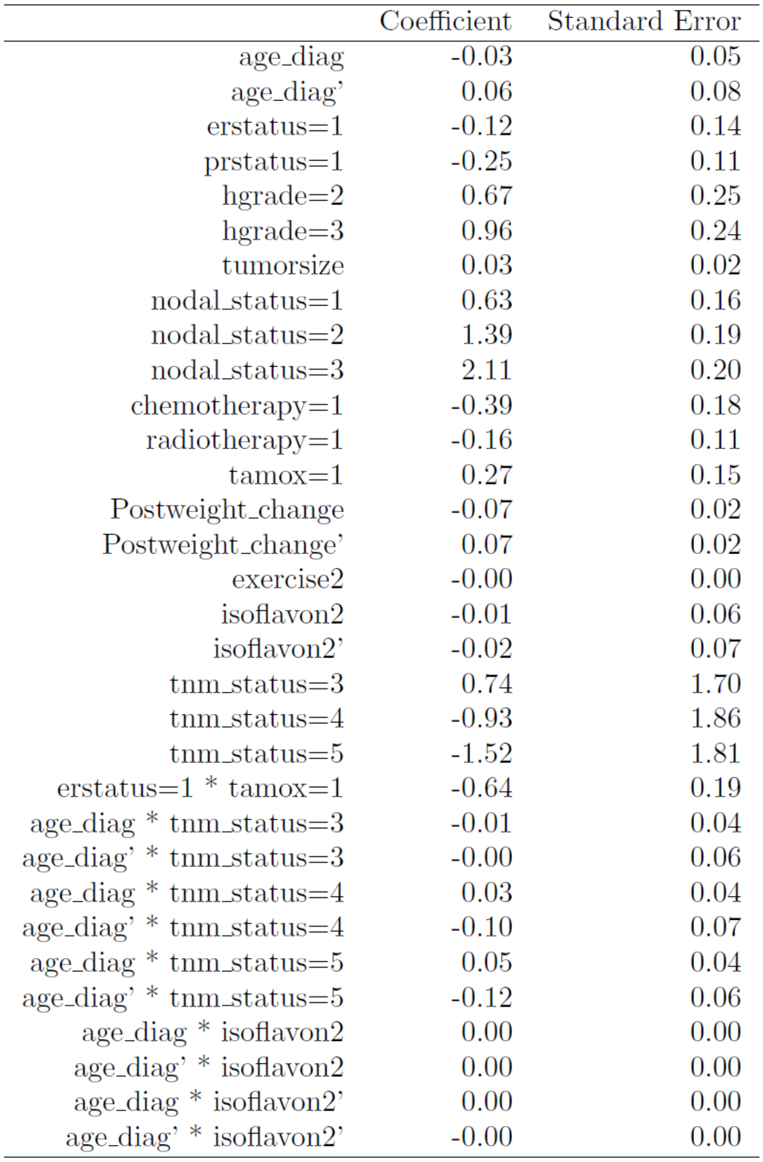
	Estimate	Std.Error	95% CI lower bound	95% CI upper bound
age_diag	-0.02	0.05	-0.12	0.07
age_diag’	0.05	0.08	-0.11	0.21
erstatus=1	-0.13	0.14	-0.39	0.14
prstatus=1	-0.25	0.11	-0.47	-0.04
her_2=1	0.03	0.19	-0.33	0.39
her_2=2	0.02	0.11	-0.18	0.23
hgrade=2	0.67	0.25	0.18	1.17
hgrade=3	0.96	0.24	0.48	1.44
tumorsize	0.03	0.02	0.00	0.06
nodal_status=1	0.64	0.16	0.33	0.95
nodal_status=2	1.42	0.19	1.05	1.78
nodal_status=3	2.14	0.20	1.75	2.53
chemotherapy=1	-0.37	0.18	-0.73	-0.01
radiotherapy=1	-0.19	0.11	-0.41	0.02
tamox=1	0.27	0.15	-0.01	0.56
bc_surgery=1	-0.20	0.22	-0.63	0.22
bmi2	0.00	0.01	-0.02	0.03
Postweight_change	-0.07	0.02	-0.11	-0.03
Postweight_change’	0.07	0.02	0.03	0.12
exercise2	-0.00	0.00	-0.01	0.00
isoflavon2	-0.01	0.06	-0.12	0.10
isoflavon2’	-0.02	0.07	-0.17	0.12
tnm_status=3	0.80	1.70	-2.54	4.13
tnm_status=4	-0.91	1.85	-4.54	2.73
tnm_status=5	-1.49	1.81	-5.04	2.05
erstatus=1 * tamox=1	-0.63	0.19	-1.01	-0.25
age_diag * tnm_status=3	-0.01	0.04	-0.08	0.06
age_diag’ * tnm_status=3	-0.00	0.06	-0.12	0.12
age_diag * tnm_status=4	0.03	0.04	-0.05	0.11
age_diag’ * tnm_status=4	-0.10	0.07	-0.22	0.03
age_diag * tnm_status=5	0.05	0.04	-0.03	0.12
age_diag’ * tnm_status=5	-0.12	0.06	-0.25	0.00
age_diag * isoflavon2	0.00	0.00	-0.00	0.00
age_diag’ * isoflavon2	0.00	0.00	-0.00	0.01
age_diag * isoflavon2’	0.00	0.00	-0.00	0.00
age_diag’ * isoflavon2’	-0.00	0.00	-0.01	0.00

Table 9: Discrimination Ability of 5-year Overall Survival Model



Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.5329	0.5481	0.4938	0.0247	0.5082	200
R^2	0.1360	0.1426	0.1201	0.0100	0.1259	200
Slope	1.0000	1.0000	0.8676	0.0837	0.9163	200
D	0.0655	0.0692	0.0615	0.0025	0.0630	200
U	−0.0002	−0.0002	0.0027	−0.0018	0.0016	200
Q	0.0657	0.0694	0.0588	0.0044	0.0614	200
g	1.0757	1.1282	0.9794	0.0608	1.0148	200

Table 10: Summary of Approximate Model of 5-year Overall Survival Model

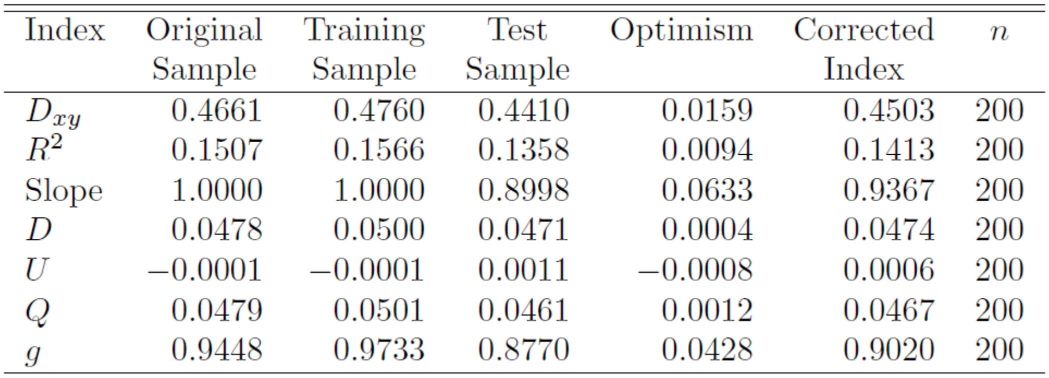


	Coefficient	Standard Error
age_diag	-0.03	0.05
age_diag'	0.06	0.08
erstatus=1	-0.12	0.14
prstatus=1	-0.25	0.11
hgrade=2	0.67	0.25
hgrade=3	0.96	0.24
tumorsize	0.03	0.02
nodal_status=1	0.63	0.16
nodal_status=2	1.39	0.19
nodal_status=3	2.11	0.20
chemotherapy=1	-0.39	0.18
radiotherapy=1	-0.16	0.11
tamox=1	0.27	0.15
Postweight_change	-0.07	0.02
Postweight_change'	0.07	0.02
exercise2	-0.00	0.00
isoflavon2	-0.01	0.06
isoflavon2'	-0.02	0.07
tnm_status=3	0.74	1.70
tnm_status=4	-0.93	1.86
tnm_status=5	-1.52	1.81
erstatus=1 * tamox=1	-0.64	0.19
age_diag * tnm_status=3	-0.01	0.04
age_diag' * tnm_status=3	-0.00	0.06
age_diag * tnm_status=4	0.03	0.04
age_diag' * tnm_status=4	-0.10	0.07
age_diag * tnm_status=5	0.05	0.04
age_diag' * tnm_status=5	-0.12	0.06
age_diag * isoflavon2	0.00	0.00
age_diag' * isoflavon2	0.00	0.00
age_diag * isoflavon2'	0.00	0.00
age_diag' * isoflavon2'	-0.00	0.00

Table 11: Summary of 10-year Overall Survival Model

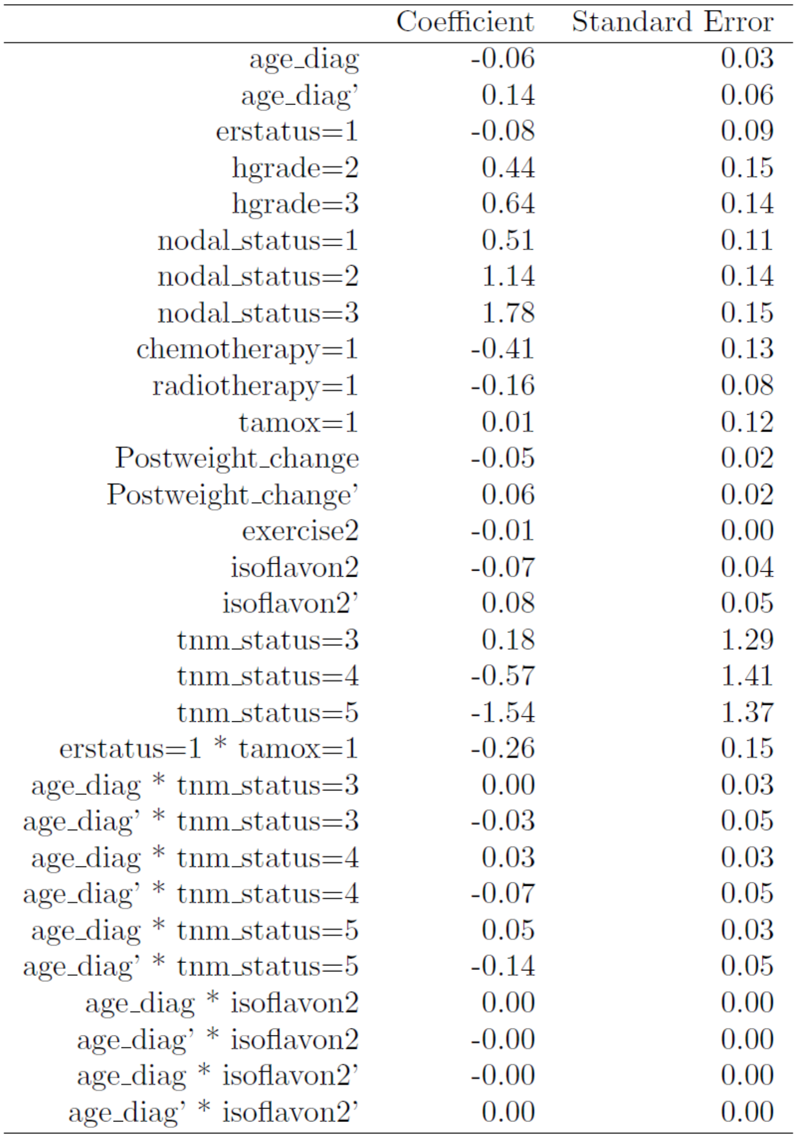
	Estimate	Std.Error	95% CI lower bound	95% CI upper bound
age_diag	-0.06	0.03	-0.12	0.00
age_diag'	0.14	0.06	0.03	0.25
erstatus=1	-0.01	0.10	-0.21	0.19
prstatus=1	-0.13	0.08	-0.30	0.03
her_2=1	-0.06	0.14	-0.34	0.22
her_2=2	0.02	0.08	-0.15	0.18
hgrade=2	0.44	0.15	0.15	0.73
hgrade=3	0.63	0.14	0.35	0.91
tumorsize	0.03	0.01	0.00	0.05
nodal_status=1	0.56	0.11	0.34	0.78
nodal_status=2	1.20	0.14	0.92	1.48
nodal_status=3	1.84	0.16	1.53	2.14
chemotherapy=1	-0.39	0.13	-0.64	-0.14
radiotherapy=1	-0.18	0.08	-0.34	-0.01
tamox=1	0.04	0.12	-0.20	0.29
bc_surgery=1	-0.12	0.16	-0.42	0.18
bmi2	0.01	0.01	-0.01	0.03
Postweight_change	-0.05	0.02	-0.08	-0.02
Postweight_change'	0.06	0.02	0.02	0.09
exercise2	-0.01	0.00	-0.01	-0.00
isoflavon2	-0.07	0.04	-0.15	0.01
isoflavon2'	0.08	0.05	-0.02	0.17
tnm_status=3	0.16	1.29	-2.36	2.68
tnm_status=4	-0.66	1.41	-3.42	2.09
tnm_status=5	-1.64	1.37	-4.32	1.04
erstatus=1 * tamox=1	-0.28	0.15	-0.57	0.02
age_diag * tnm_status=3	0.00	0.03	-0.05	0.06
age_diag' * tnm_status=3	-0.03	0.05	-0.12	0.06
age_diag * tnm_status=4	0.03	0.03	-0.03	0.08
age_diag' * tnm_status=4	-0.08	0.05	-0.17	0.02
age_diag * tnm_status=5	0.05	0.03	-0.00	0.11
age_diag' * tnm_status=5	-0.14	0.05	-0.23	-0.04
age_diag * isoflavon2	0.00	0.00	-0.00	0.00
age_diag' * isoflavon2	-0.00	0.00	-0.00	0.00
age_diag * isoflavon2'	-0.00	0.00	-0.00	0.00
age_diag' * isoflavon2'	0.00	0.00	-0.00	0.01

Table 12: Discrimination Ability of 10-year Overall Survival Model



Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.4661	0.4760	0.4410	0.0159	0.4503	200
R^2	0.1507	0.1566	0.1358	0.0094	0.1413	200
Slope	1.0000	1.0000	0.8998	0.0633	0.9367	200
D	0.0478	0.0500	0.0471	0.0004	0.0474	200
U	−0.0001	−0.0001	0.0011	−0.0008	0.0006	200
Q	0.0479	0.0501	0.0461	0.0012	0.0467	200
g	0.9448	0.9733	0.8770	0.0428	0.9020	200

Table 13: Summary of Approximate Model of 10-year Overall Survival Model



	Coefficient	Standard Error
age_diag	-0.06	0.03
age_diag’	0.14	0.06
erstatus=1	-0.08	0.09
hgrade=2	0.44	0.15
hgrade=3	0.64	0.14
nodal_status=1	0.51	0.11
nodal_status=2	1.14	0.14
nodal_status=3	1.78	0.15
chemotherapy=1	-0.41	0.13
radiotherapy=1	-0.16	0.08
tamox=1	0.01	0.12
Postweight_change	-0.05	0.02
Postweight_change’	0.06	0.02
exercise2	-0.01	0.00
isoflavon2	-0.07	0.04
isoflavon2’	0.08	0.05
tnm_status=3	0.18	1.29
tnm_status=4	-0.57	1.41
tnm_status=5	-1.54	1.37
erstatus=1 * tamox=1	-0.26	0.15
age_diag * tnm_status=3	0.00	0.03
age_diag’ * tnm_status=3	-0.03	0.05
age_diag * tnm_status=4	0.03	0.03
age_diag’ * tnm_status=4	-0.07	0.05
age_diag * tnm_status=5	0.05	0.03
age_diag’ * tnm_status=5	-0.14	0.05
age_diag * isoflavon2	0.00	0.00
age_diag’ * isoflavon2	-0.00	0.00
age_diag * isoflavon2’	-0.00	0.00
age_diag’ * isoflavon2’	0.00	0.00

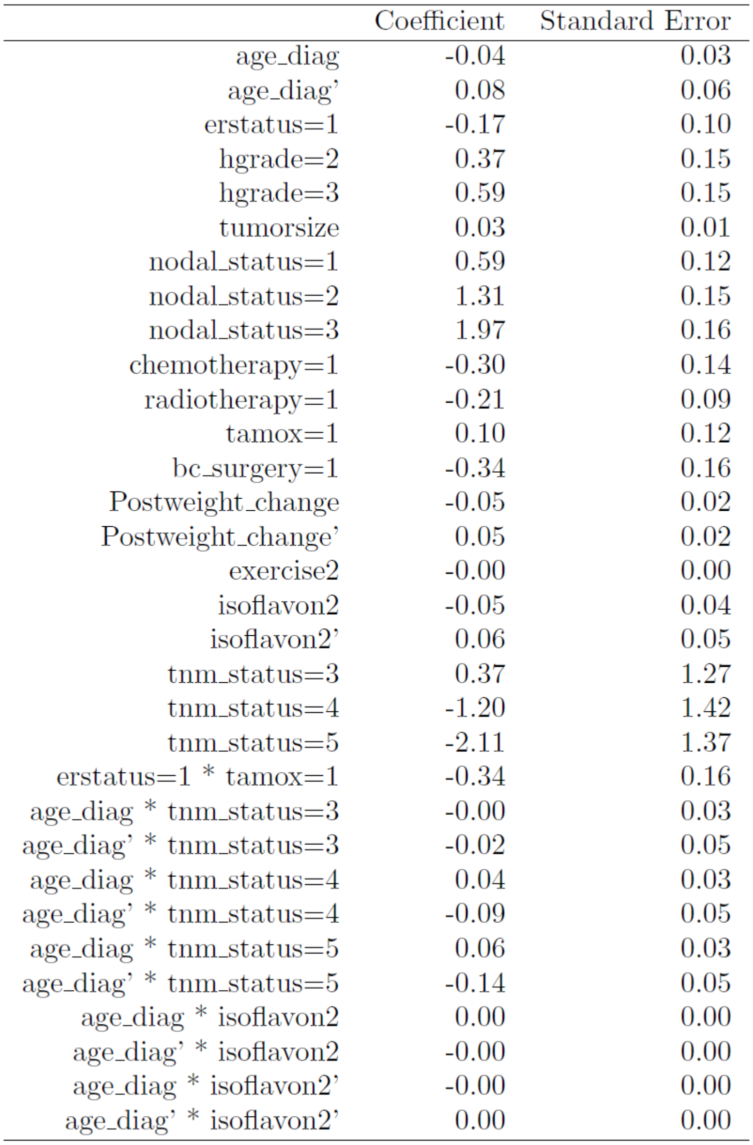
Table 14: Summary of 5-year Relapse-free Survival Model

	Estimate	Std.Error	95% CI lower bound	95% CI upper bound
age_diag	-0.04	0.03	-0.11	0.02
age_diag'	0.09	0.05	-0.02	0.19
erstatus=1	-0.12	0.11	-0.34	0.10
prstatus=1	-0.09	0.09	-0.27	0.09
her_2=1	0.03	0.15	-0.27	0.32
her_2=2	0.02	0.09	-0.15	0.20
hgrade=2	0.37	0.15	0.08	0.67
hgrade=3	0.57	0.15	0.28	0.87
tumorsize	0.03	0.01	0.00	0.05
nodal_status=1	0.59	0.12	0.35	0.83
nodal_status=2	1.31	0.15	1.02	1.61
nodal_status=3	1.97	0.16	1.65	2.29
chemotherapy=1	-0.30	0.14	-0.58	-0.02
radiotherapy=1	-0.21	0.09	-0.38	-0.04
tamox=1	0.12	0.13	-0.12	0.37
bc_surgery=1	-0.35	0.16	-0.66	-0.04
bmi2	0.01	0.01	-0.01	0.03
Postweight_change	-0.05	0.02	-0.08	-0.02
Postweight_change'	0.05	0.02	0.01	0.09
exercise2	-0.00	0.00	-0.01	0.00
isoflavon2	-0.05	0.04	-0.12	0.03
isoflavon2'	0.06	0.05	-0.04	0.16
tnm_status=3	0.36	1.27	-2.13	2.86
tnm_status=4	-1.23	1.42	-4.01	1.54
tnm_status=5	-2.12	1.36	-4.80	0.55
erstatus=1 * tamox=1	-0.36	0.16	-0.66	-0.05
age_diag * tnm_status=3	-0.00	0.03	-0.05	0.05
age_diag' * tnm_status=3	-0.02	0.05	-0.12	0.07
age_diag * tnm_status=4	0.04	0.03	-0.02	0.10
age_diag' * tnm_status=4	-0.09	0.05	-0.19	0.01
age_diag * tnm_status=5	0.06	0.03	0.00	0.12
age_diag' * tnm_status=5	-0.14	0.05	-0.24	-0.05
age_diag * isoflavon2	0.00	0.00	-0.00	0.00
age_diag' * isoflavon2	-0.00	0.00	-0.00	0.00
age_diag * isoflavon2'	-0.00	0.00	-0.00	0.00
age_diag' * isoflavon2'	0.00	0.00	-0.00	0.00

Table 15: Discrimination Ability of 5-year Relapse-free Survival Model

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.4533	0.4638	0.4225	0.0194	0.4338	200
R^2	0.1340	0.1392	0.1174	0.0105	0.1235	200
Slope	1.0000	1.0000	0.8837	0.0735	0.9265	200
D	0.0464	0.0485	0.0443	0.0013	0.0450	200
U	−0.0001	−0.0001	0.0017	−0.0011	0.0010	200
Q	0.0465	0.0486	0.0426	0.0024	0.0441	200
g	0.8935	0.9373	0.8284	0.0412	0.8524	200

Table 16: Summary of Approximate Model of 5-year Relapse-free Survival Model



	Coefficient	Standard Error
age_diag	-0.04	0.03
age_diag’	0.08	0.06
erstatus=1	-0.17	0.10
hgrade=2	0.37	0.15
hgrade=3	0.59	0.15
tumorsize	0.03	0.01
nodal_status=1	0.59	0.12
nodal_status=2	1.31	0.15
nodal_status=3	1.97	0.16
chemotherapy=1	-0.30	0.14
radiotherapy=1	-0.21	0.09
tamox=1	0.10	0.12
bc_surgery=1	-0.34	0.16
Postweight_change	-0.05	0.02
Postweight_change’	0.05	0.02
exercise2	-0.00	0.00
isoflavon2	-0.05	0.04
isoflavon2’	0.06	0.05
tnm_status=3	0.37	1.27
tnm_status=4	-1.20	1.42
tnm_status=5	-2.11	1.37
erstatus=1 * tamox=1	-0.34	0.16
age_diag * tnm_status=3	-0.00	0.03
age_diag’ * tnm_status=3	-0.02	0.05
age_diag * tnm_status=4	0.04	0.03
age_diag’ * tnm_status=4	-0.09	0.05
age_diag * tnm_status=5	0.06	0.03
age_diag’ * tnm_status=5	-0.14	0.05
age_diag * isoflavon2	0.00	0.00
age_diag’ * isoflavon2	-0.00	0.00
age_diag * isoflavon2’	-0.00	0.00
age_diag’ * isoflavon2’	0.00	0.00

Figure 1. Hierarchical Cluster Analysis of All Predictors

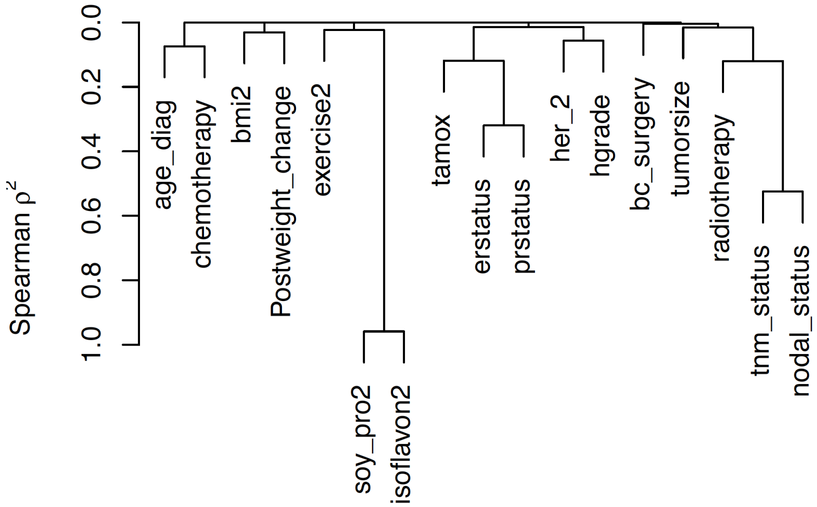


Figure 2: Fraction of NAs in each variable

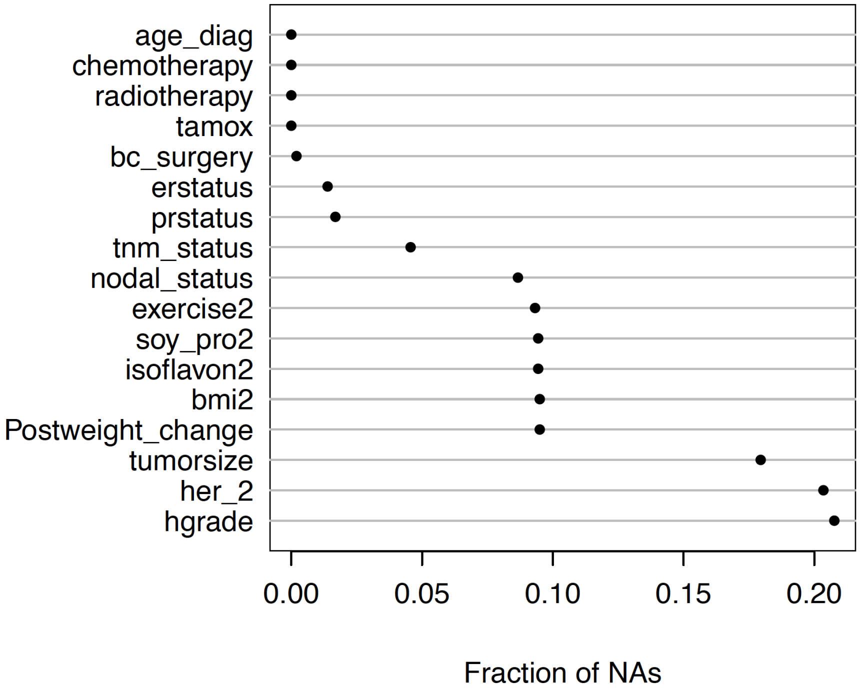


Figure 3: Variables Tend to Miss Together

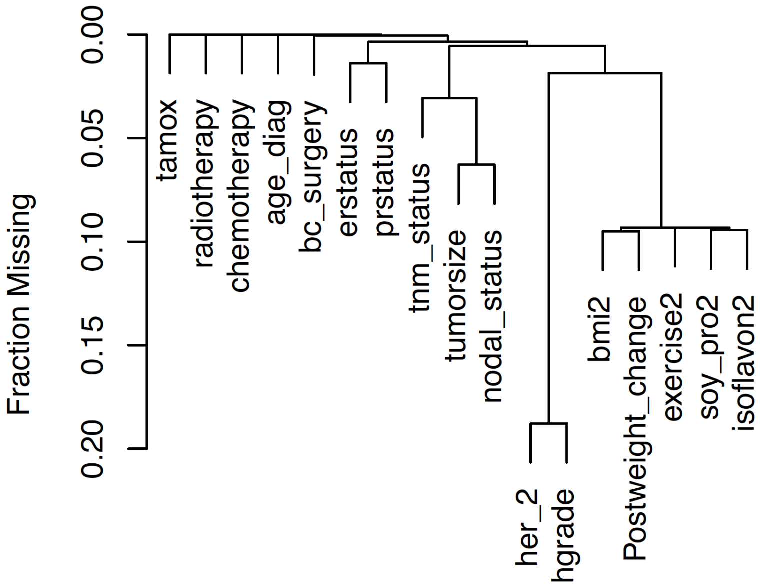
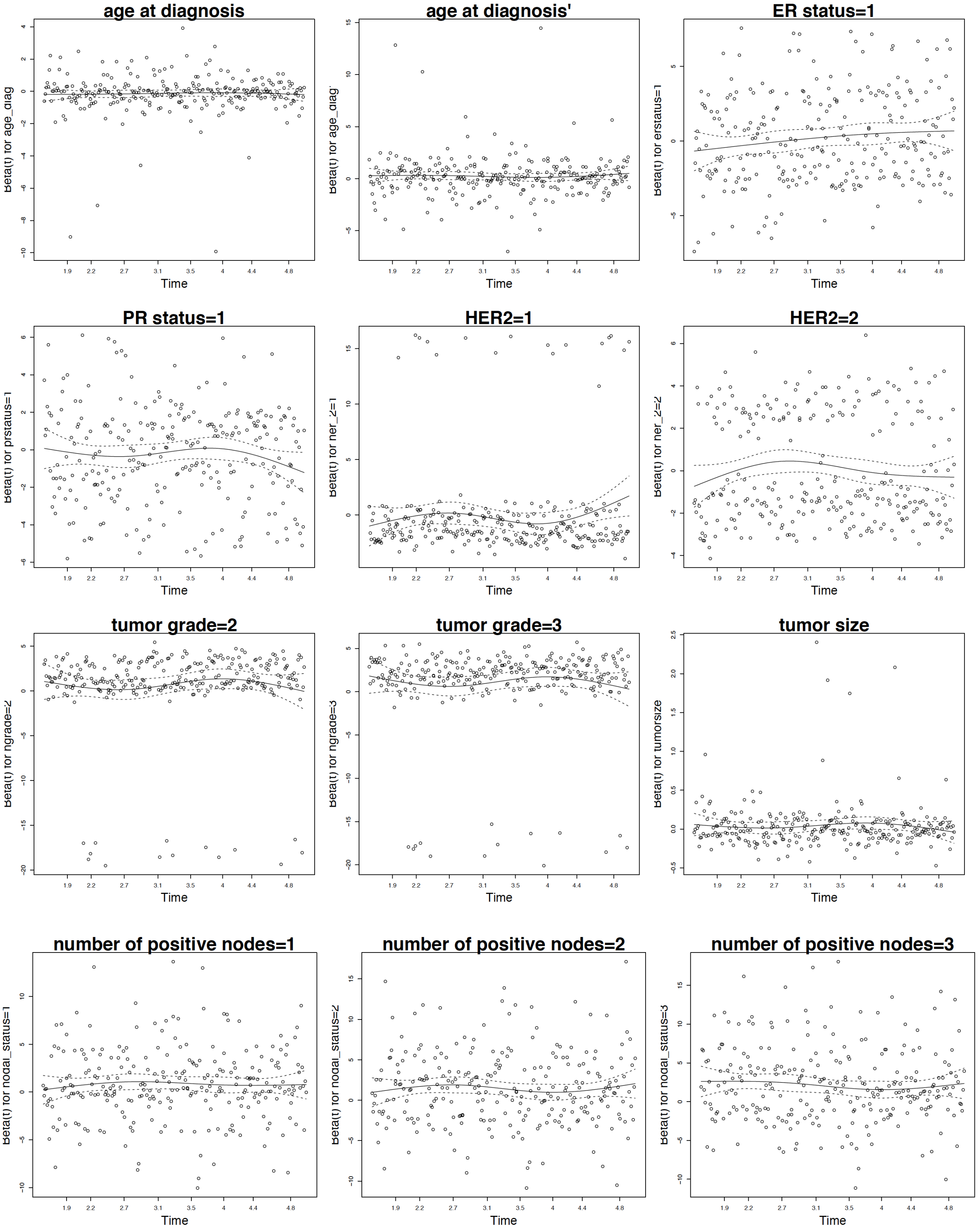
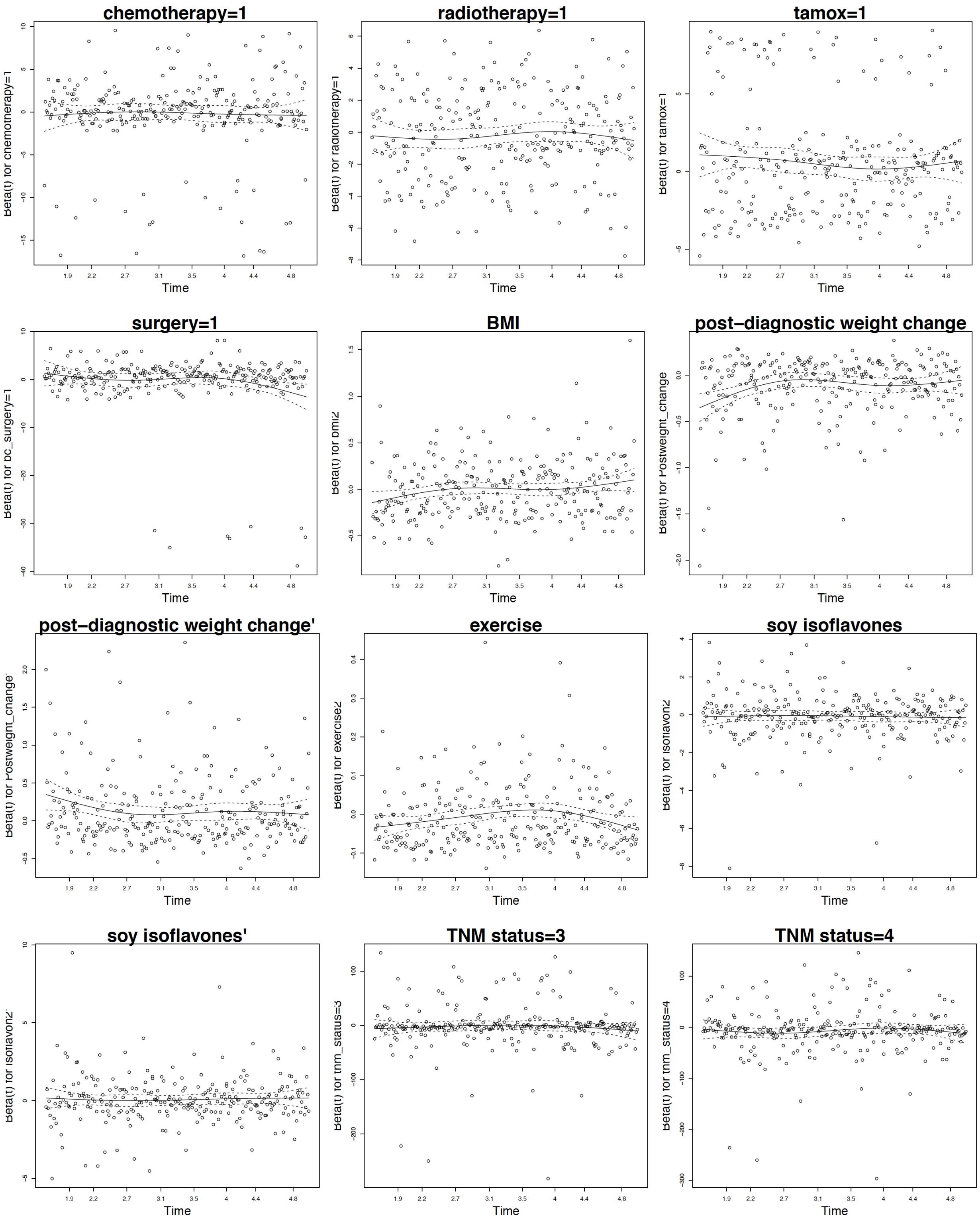


Figure 4: Schoenfeld Residuals of Individual Predictors in 5-year Overall Survival Model




chemotherapy=1

radiotherapy=1

tamox=1

surgery=1

BMI

post–diagnostic weight change

post–diagnostic weight change'

exercise

soy isoflavones

soy isoflavones'

TNM status=3

TNM status=4

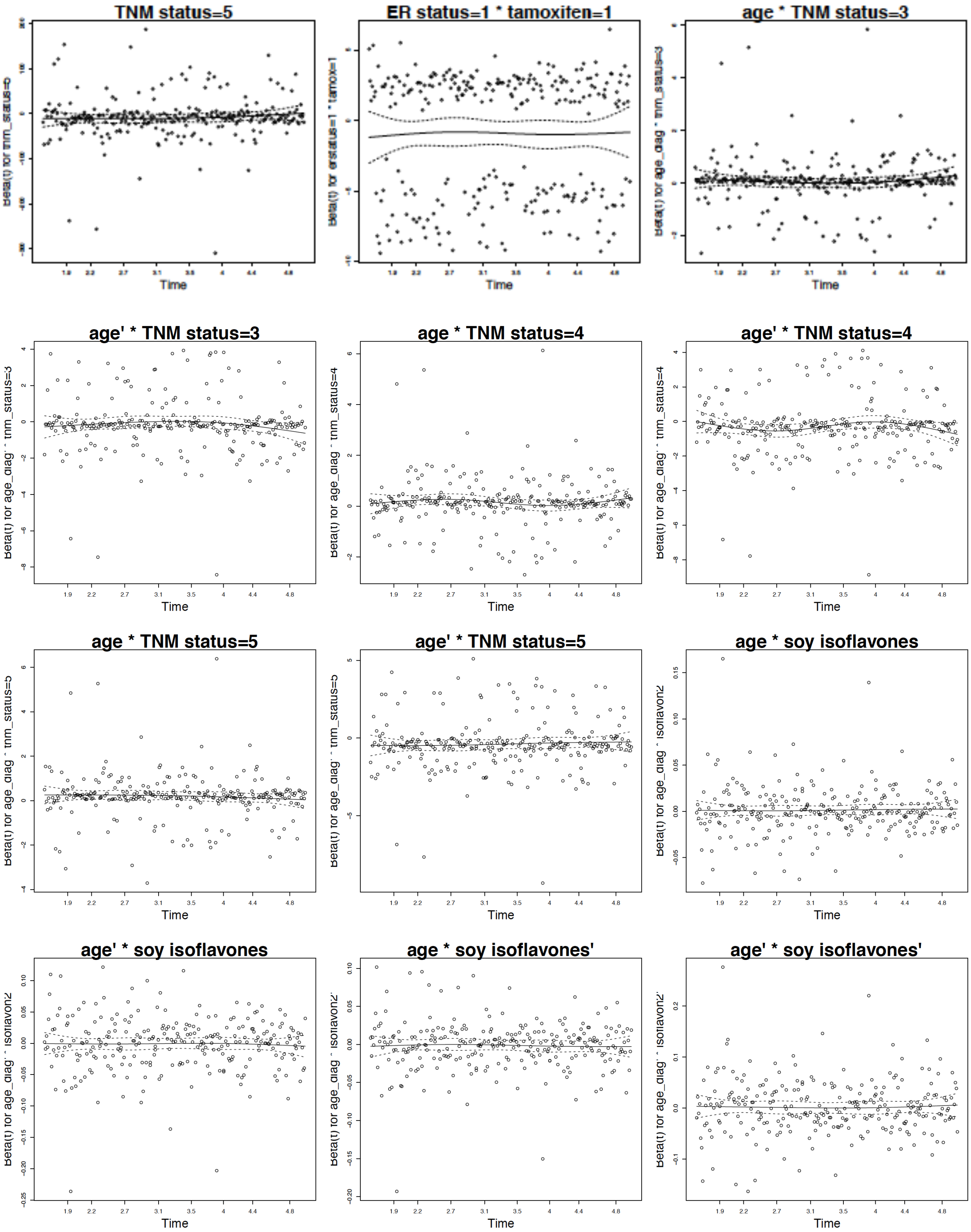

TNM status=5
Beta(t) for tnm_status=5
ER status=1 * tamoxifen=1
Beta(t) for erstatus=1 * tamox=1
age * TNM status=3
Beta(t) for age_diag * tnm_status=3
age' * TNM status=3
Beta(t) for age_diag' * tnm_status=3
age * TNM status=4
Beta(t) for age_diag * tnm_status=4
age' * TNM status=4
Beta(t) for age_diag' * tnm_status=4
age * TNM status=5
Beta(t) for age_diag * tnm_status=5
age' * TNM status=5
Beta(t) for age_diag' * tnm_status=5
age * soy isoflavones
Beta(t) for age_diag * isoflavon2
age' * soy isoflavones
Beta(t) for age_diag' * isoflavon2
age * soy isoflavones'
Beta(t) for age_diag * isoflavon2'
age' * soy isoflavones'
Beta(t) for age_diag' * isoflavon2'
Time
1.9
2.2
2.7
3.1
3.5
4
4.4
4.8

Figure 5. Nomogram of Predicting Median and Mean Survival Time in 5-year Overall Survival Model

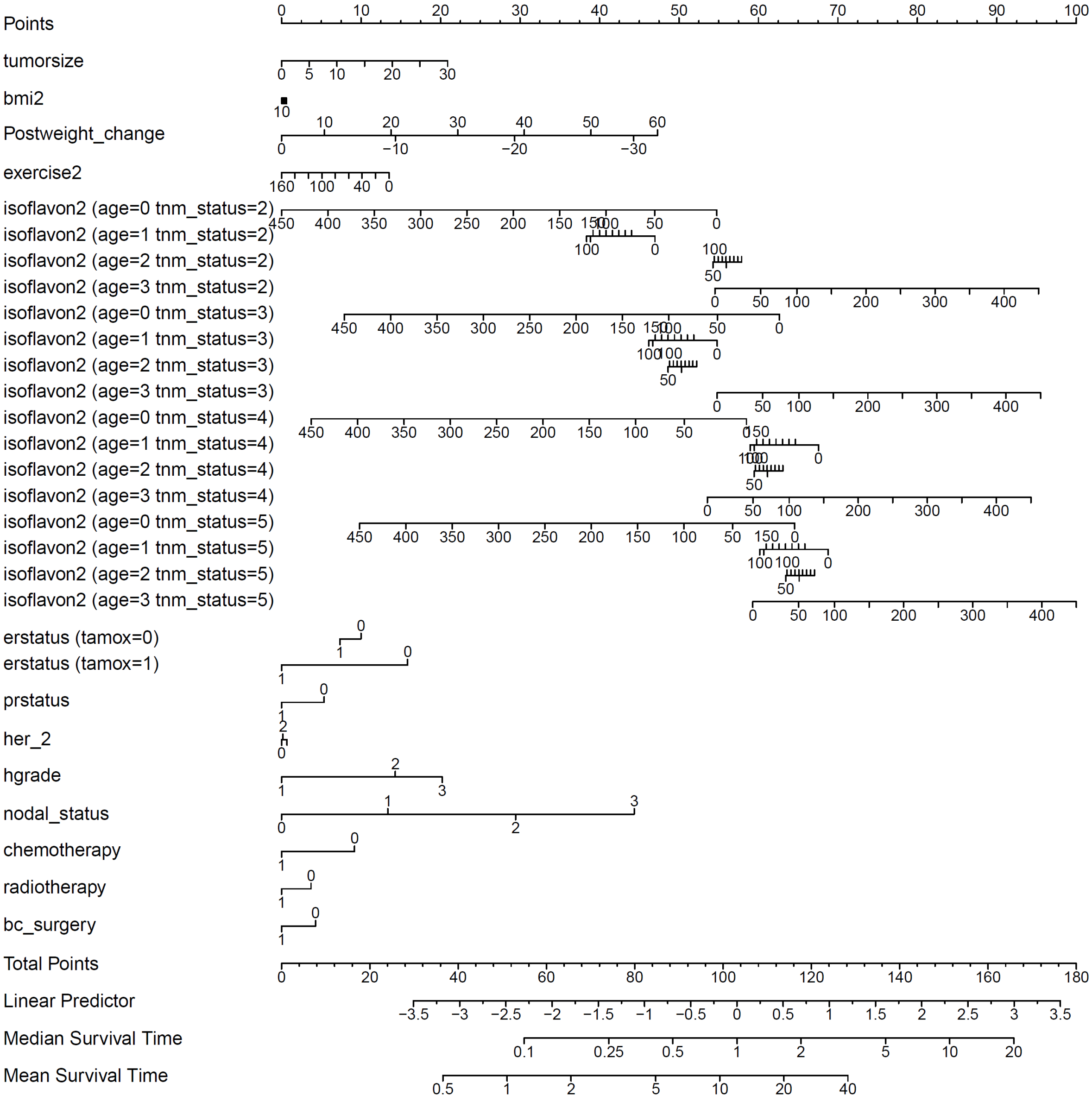


Figure 6: Contribution of Each Variable in 5-year Overall Survival Model

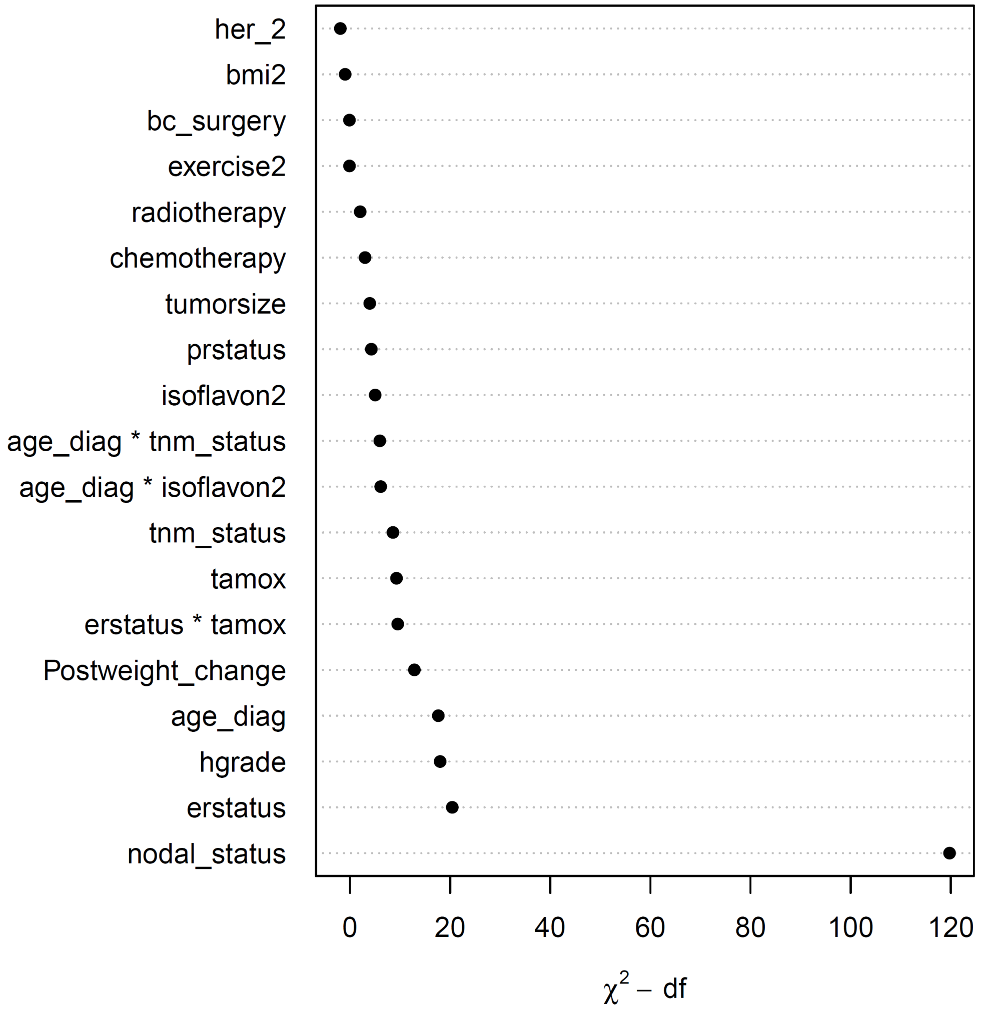


Figure 7: Estimated Hazard Ratios in 5-year Overall Survival Model

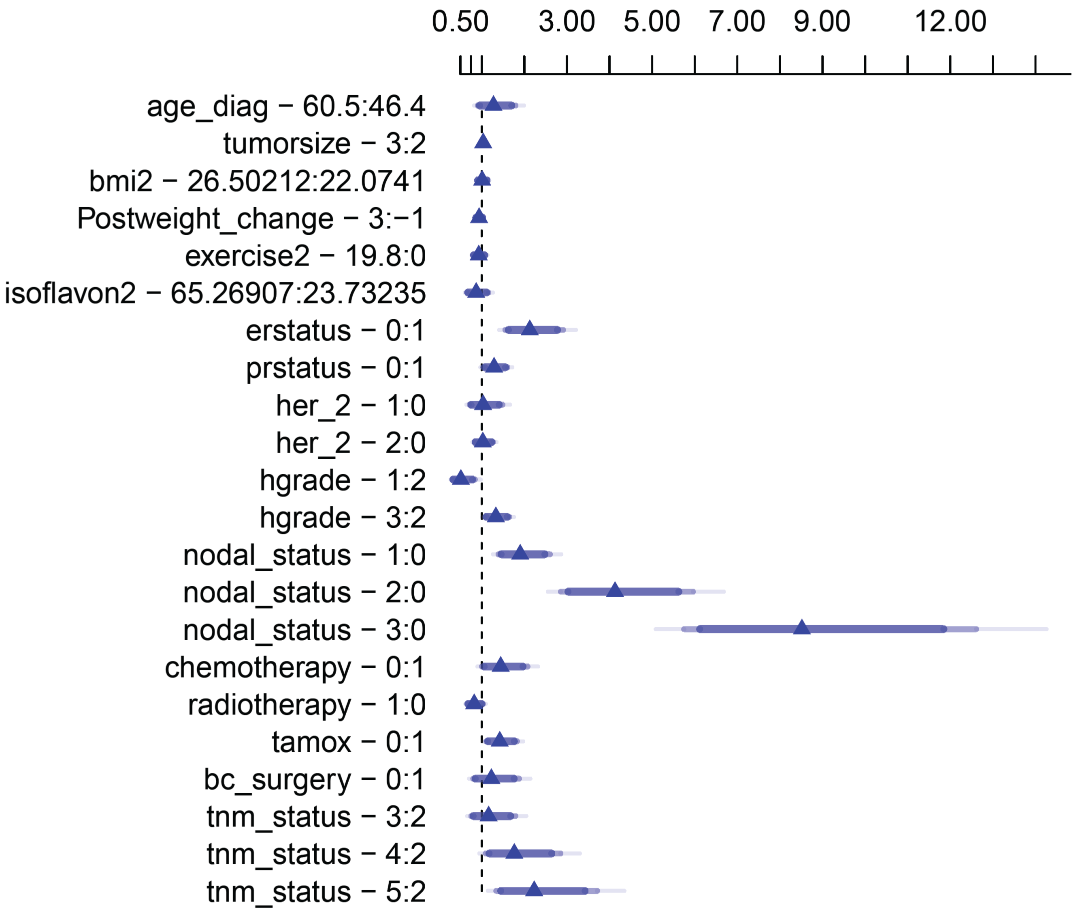


Figure 8: Effect of Each Predictor in 5-year Overall Survival Model

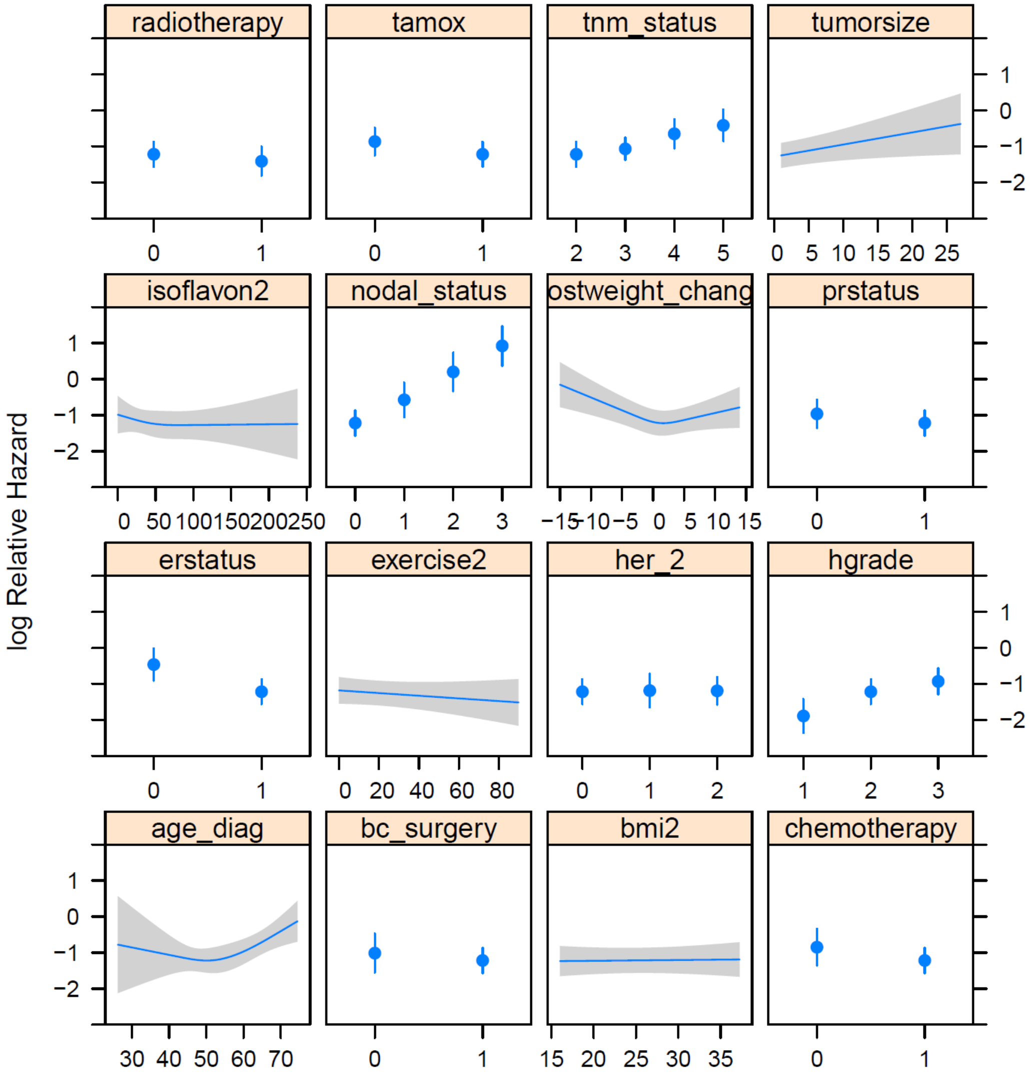


Figure 9: Effect of the Interaction of Age and Tumor-Node-Metastatic Status (TNM) in 5-year Overall Survival Model

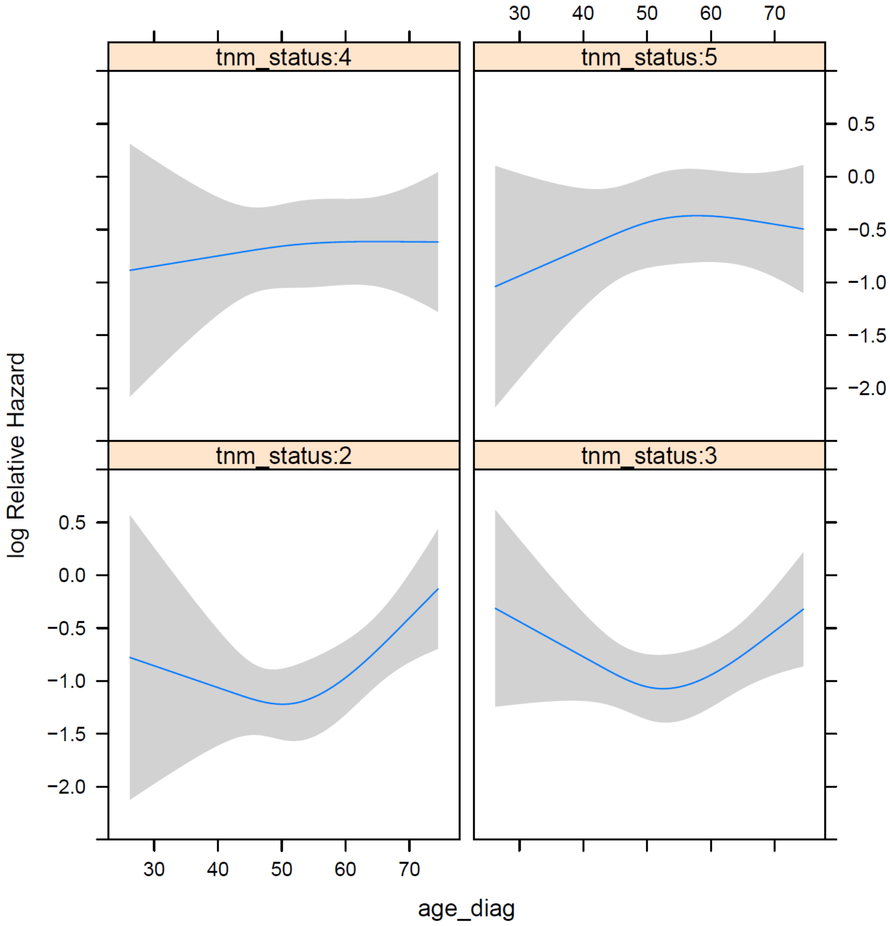


Figure 10: Effect of the Interaction of ER Status and Tamoxifen Therapy in 5-year Overall Survival Model

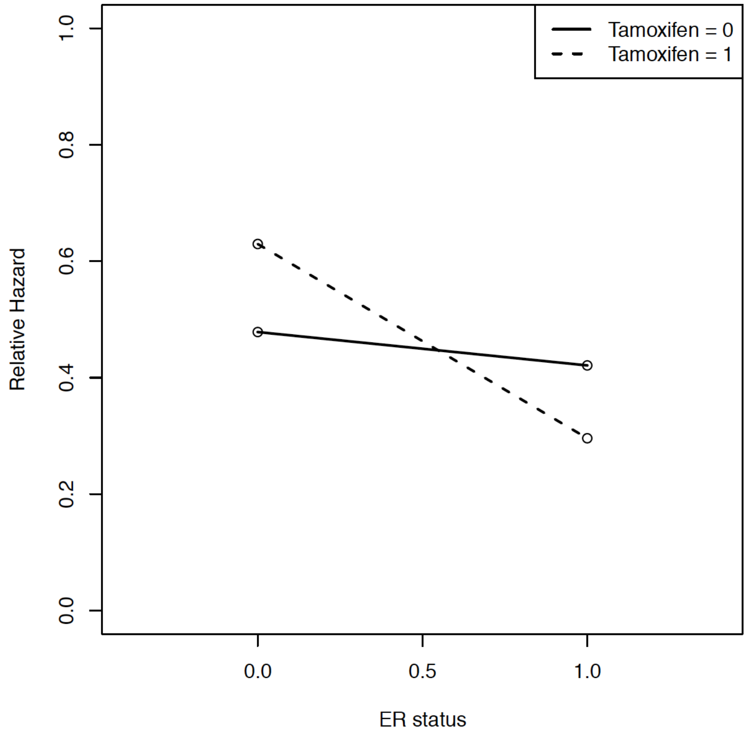


Figure 11: Effect of the Interaction of Age and Isoflavones in 5-year Overall Survival Model

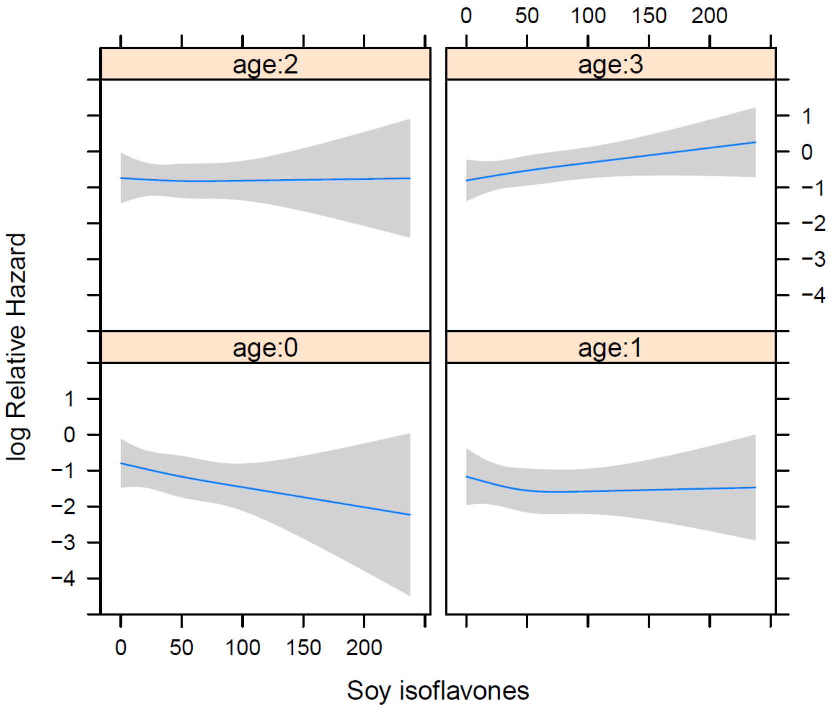


Figure 12: Survival Curve of 5-year Overall Survival Model

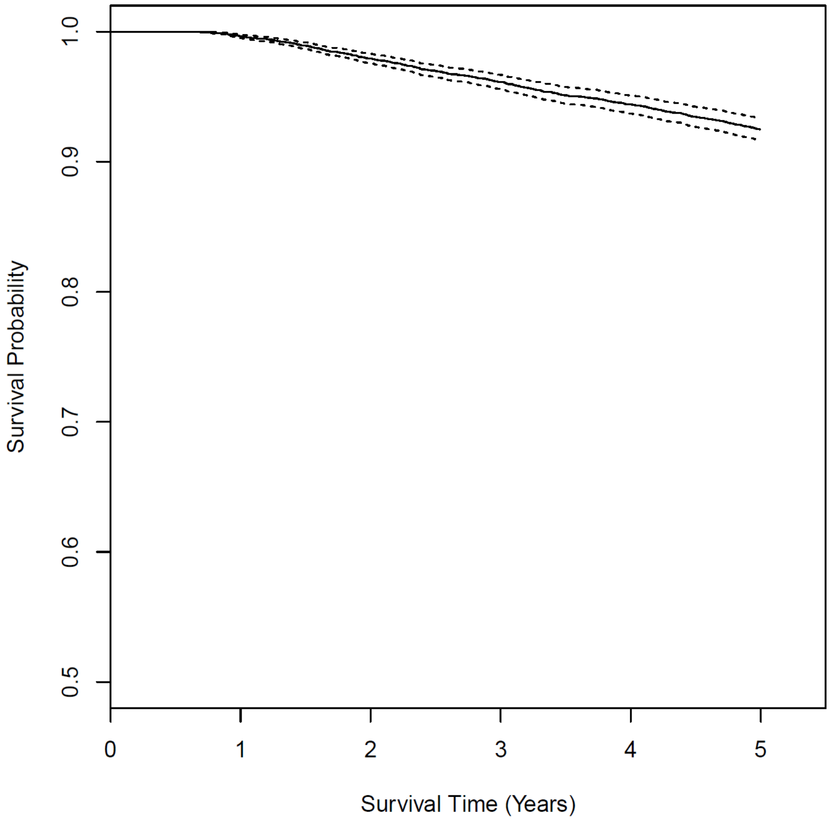


Figure 13: Calibration of 5-year Overall Survival Model

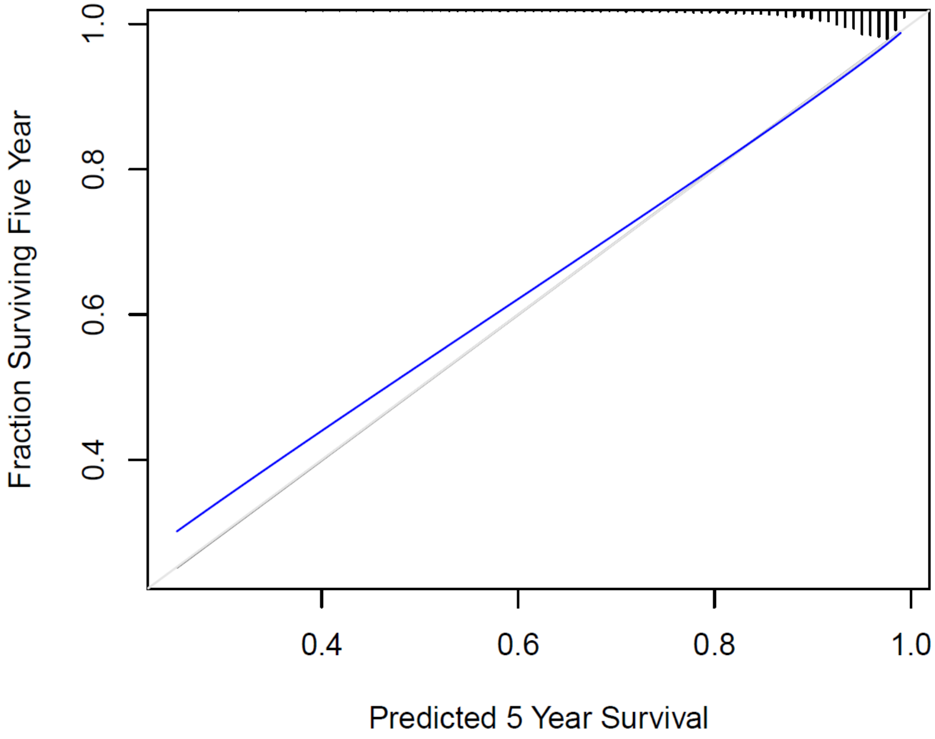
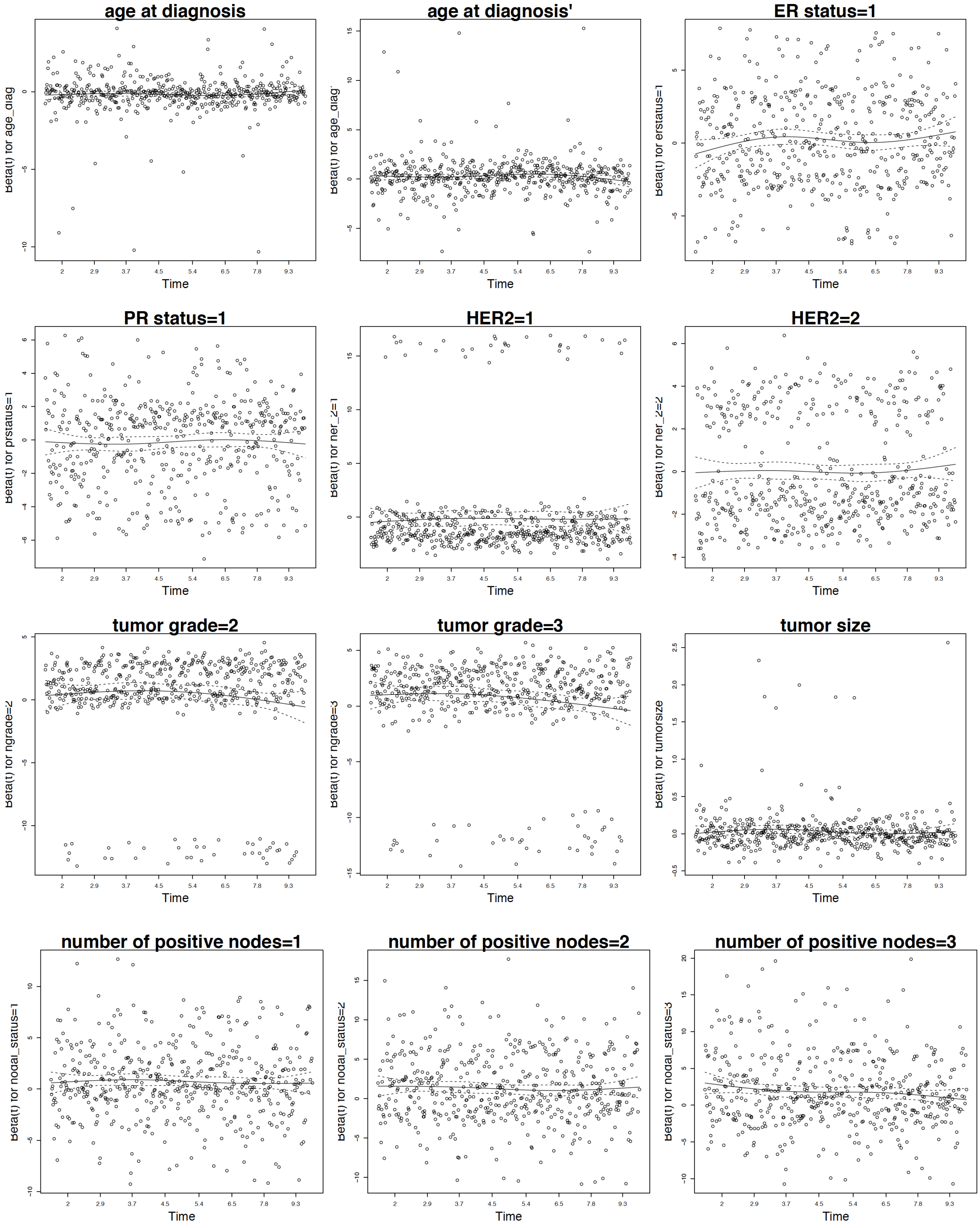
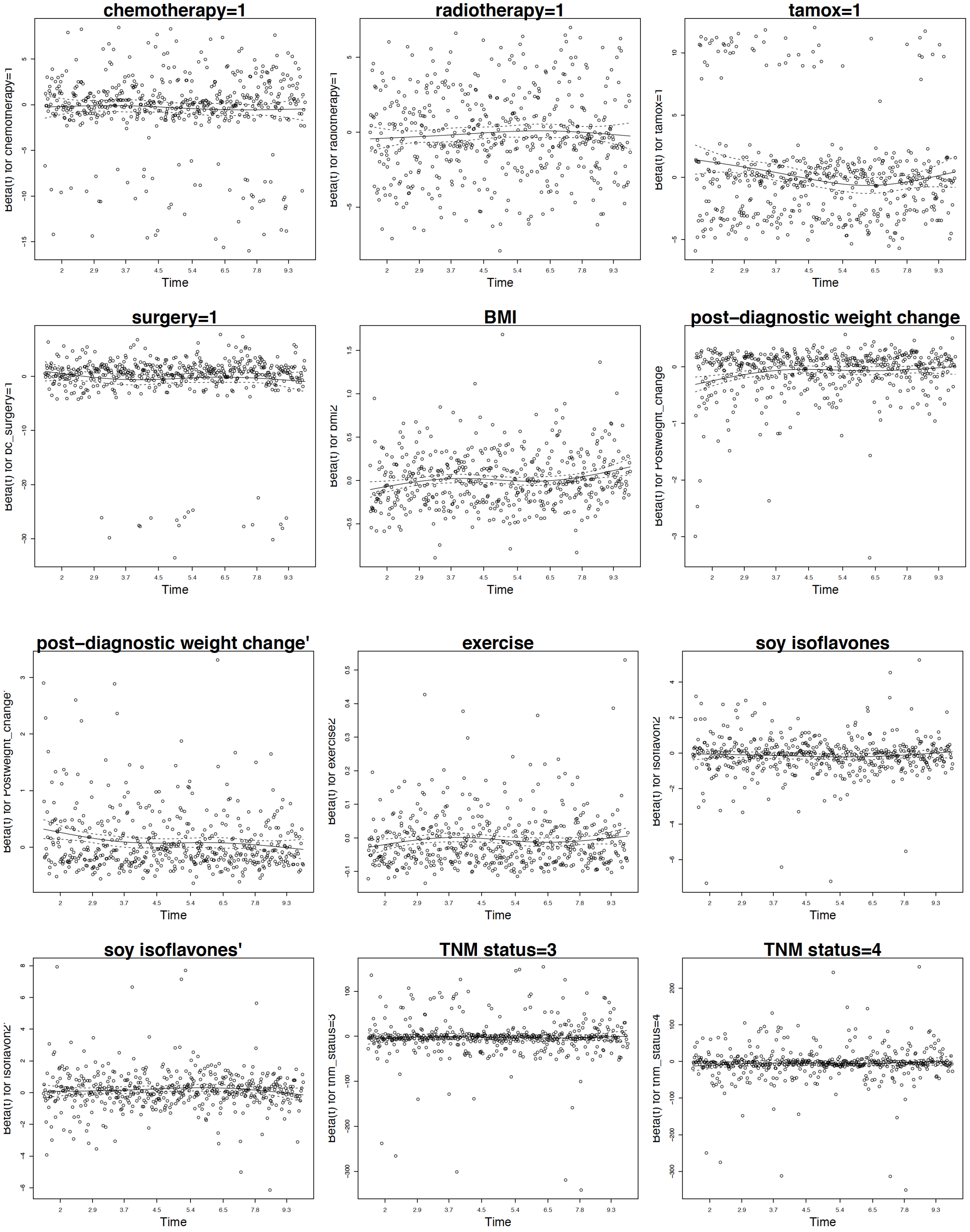


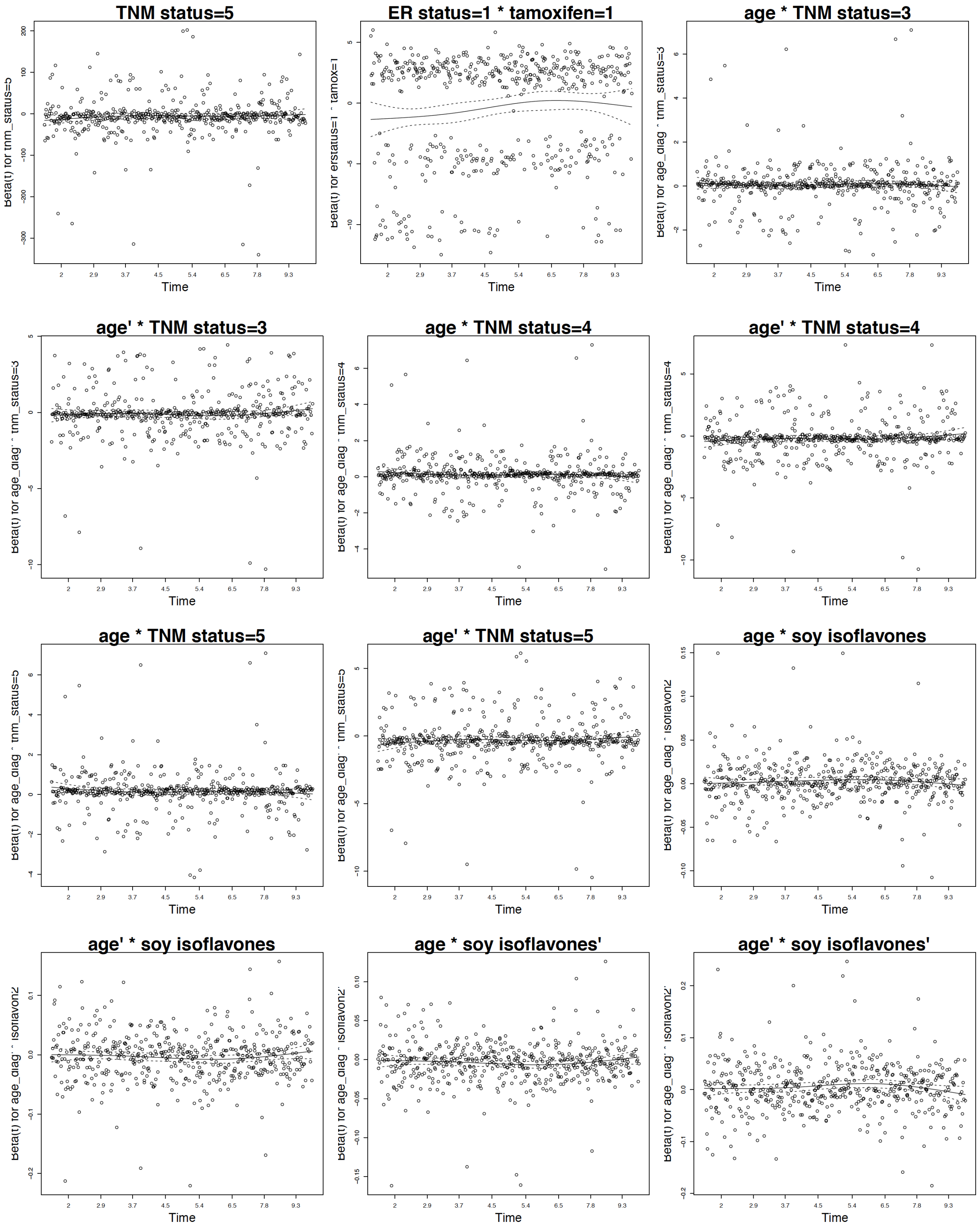
Figure 14: Factors Retained in Approximate Model of 5-year Overall Survival Model

	Demographic Predictors			Pathological Predictors			Clinical Predictors				Treatment Predictors				Lifestyle Predictors		Interaction terms		
	age at diagnosis	BMI	Post-diagnostic weight change	ER status	PR status	HER2 status	Tumor grade	Tumor size	# of Nodes	Tumor-Node -Metastasis (TNM)	Chemotherapy	Radiotherapy	Tamoxifen	Surgery	Exercise	Isoflavones	age * TNM	ER * Tamoxifen	age * isoflavones
Imputation #1	●		●	●	●		●	●	●	●	●		●			●	●	●	●
Imputation #2	●		●	●	●		●	●	●	●	●	●	●		●	●	●	●	●
Imputation #3	●		●	●	●		●	●	●	●	●	●	●		●	●	●	●	●
Imputation #4	●		●	●	●		●	●	●	●	●		●			●	●	●	●
Imputation #5	●		●	●	●		●	●	●	●	●		●		●	●	●	●	●
Imputation #6	●		●	●	●		●	●	●	●	●	●	●	●		●	●	●	●
Imputation #7	●		●	●	●		●	●	●	●	●	●	●		●	●	●	●	●
Imputation #8	●		●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●
Imputation #9	●		●	●	●		●		●	●	●	●	●			●	●	●	●
Imputation #10	●		●	●	●		●		●	●	●	●	●			●	●	●	●
Count	10	0	10	10	10	0	10	8	10	10	10	7	10	2	5	10	10	10	10

Figure 15: Schoenfeld Residuals of Individual Predictors in 10-year Overall Survival Model




chemotherapy=1
Beta(t) for chemotherapy=1
radiotherapy=1
Beta(t) for radiotherapy=1
tamox=1
Beta(t) for tamox=1
surgery=1
Beta(t) for bc_surgery=1
BMI
Beta(t) for bmi2
post−diagnostic weight change
Beta(t) for Postweight_change
post−diagnostic weight change'
Beta(t) for Postweight_change'
exercise
Beta(t) for exercise2
soy isoflavones
Beta(t) for isoflavon2
soy isoflavones'
Beta(t) for isoflavon2'
TNM status=3
Beta(t) for tnm_status=3
TNM status=4
Beta(t) for tnm_status=4
Time
2 2.9 3.7 4.5 5.4 6.5 7.8 9.3



TNM status=5

ER status=1 * tamoxifen=1

age * TNM status=3

age' * TNM status=3

age * TNM status=4

age' * TNM status=4

age * TNM status=5

age' * TNM status=5

age * soy isoflavones

age' * soy isoflavones

age * soy isoflavones'

age' * soy isoflavones'

Figure 16. Nomogram of Predicting Median and Mean Survival Time in 10-year Overall Survival Model

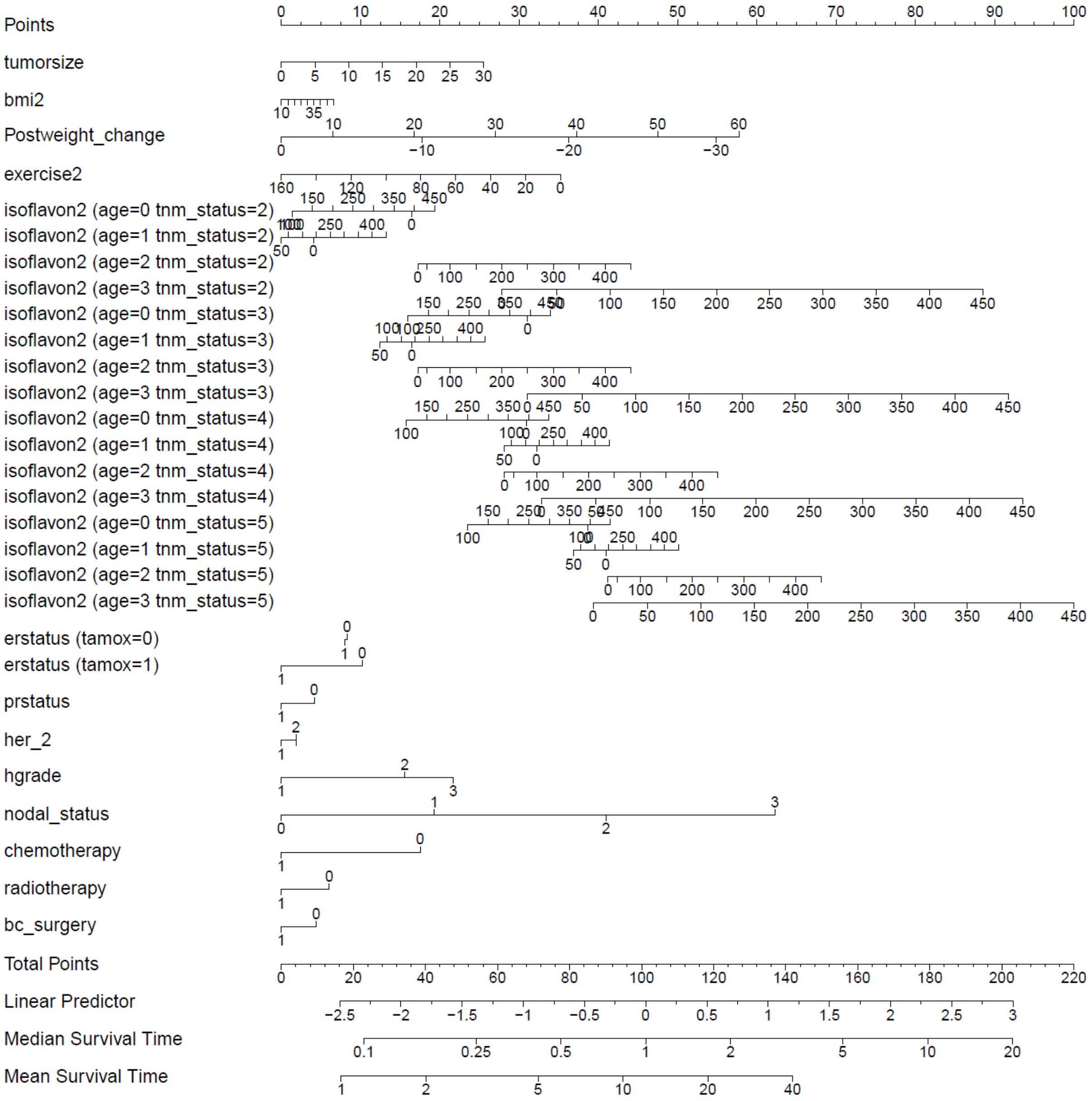


Figure 17: Contribution of Each Variable in 10-year Overall Survival Model

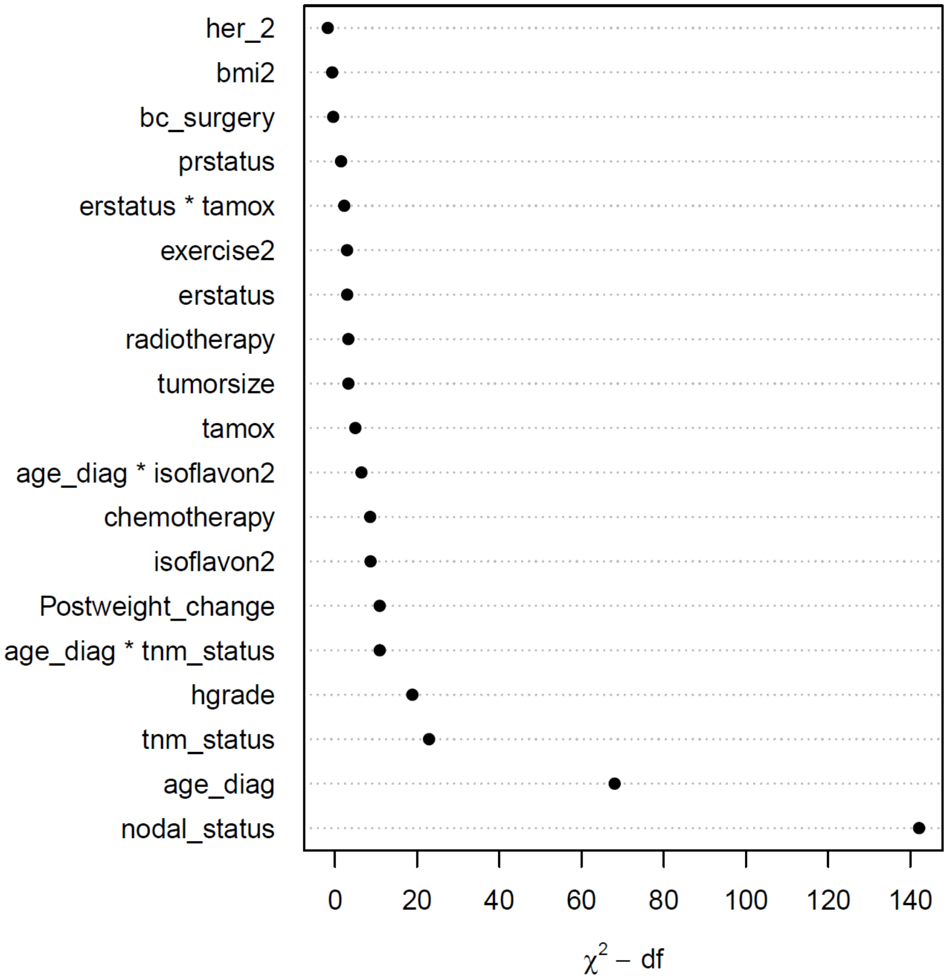


Figure 18: Estimated Hazard Ratios in 10-year Overall Survival Model

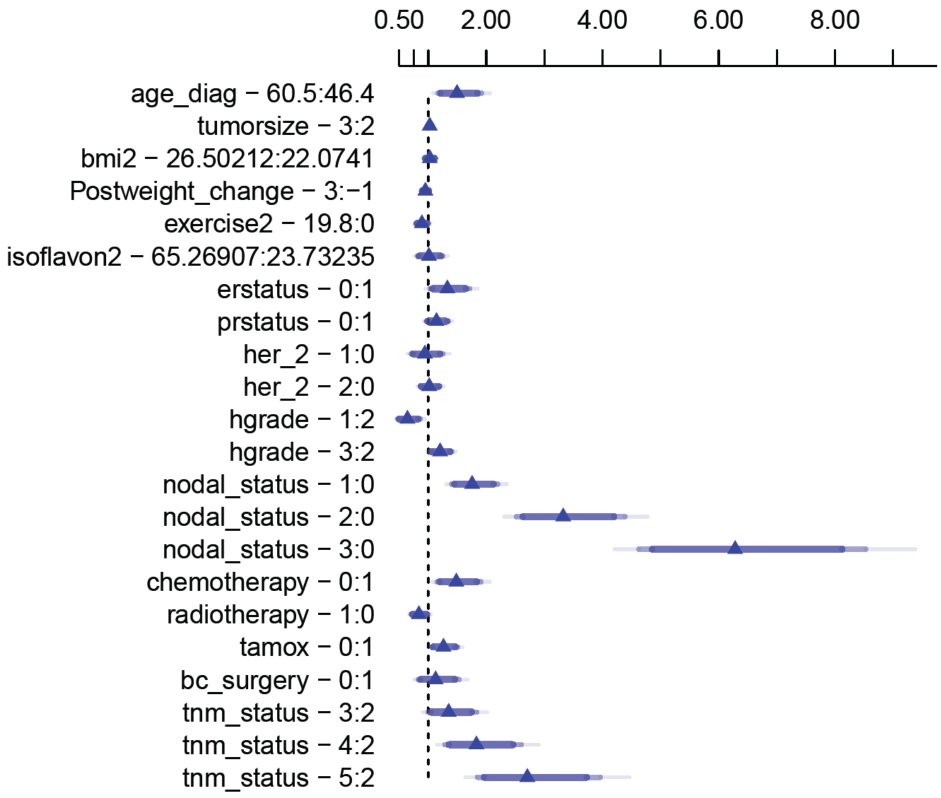


Figure 19: Effect of Each Predictor in 10-year Overall Survival Model

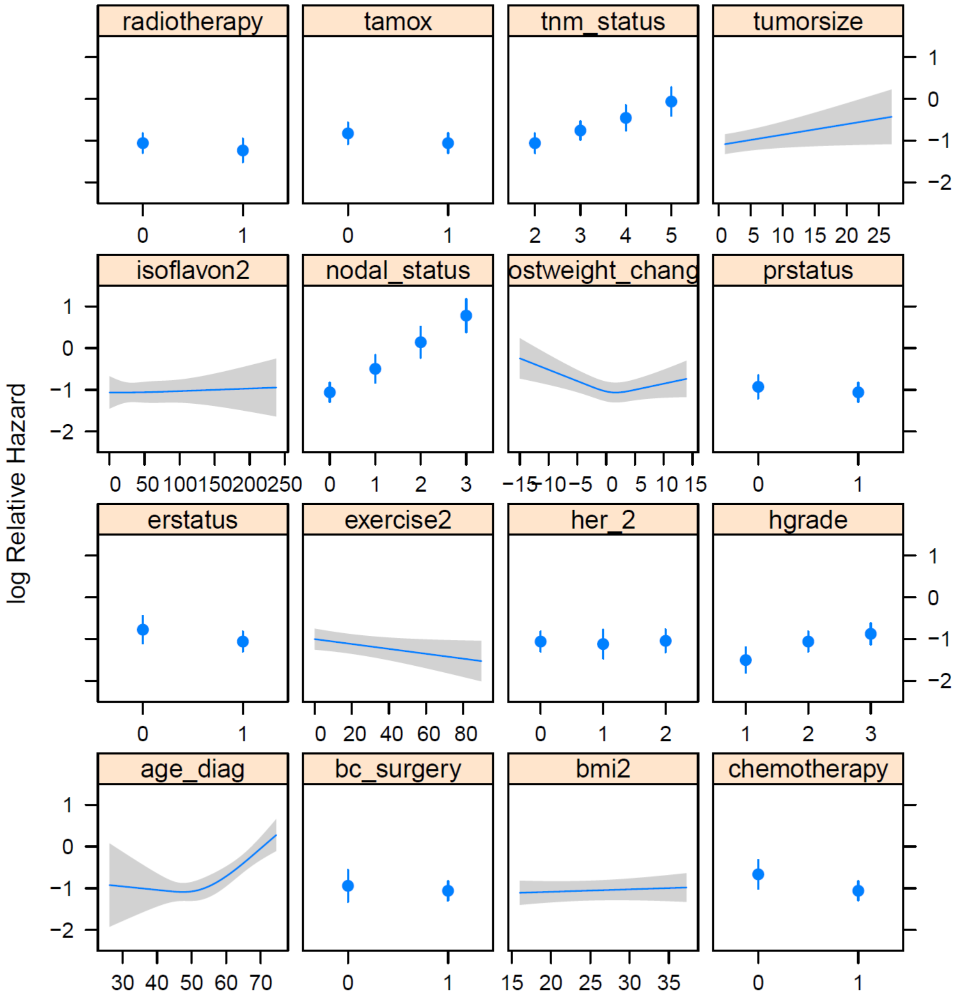


Figure 20: Effect of the Interaction of Age and Tumor-Node-Metastatic Status (TNM) in 10-yr Overall Survival Model

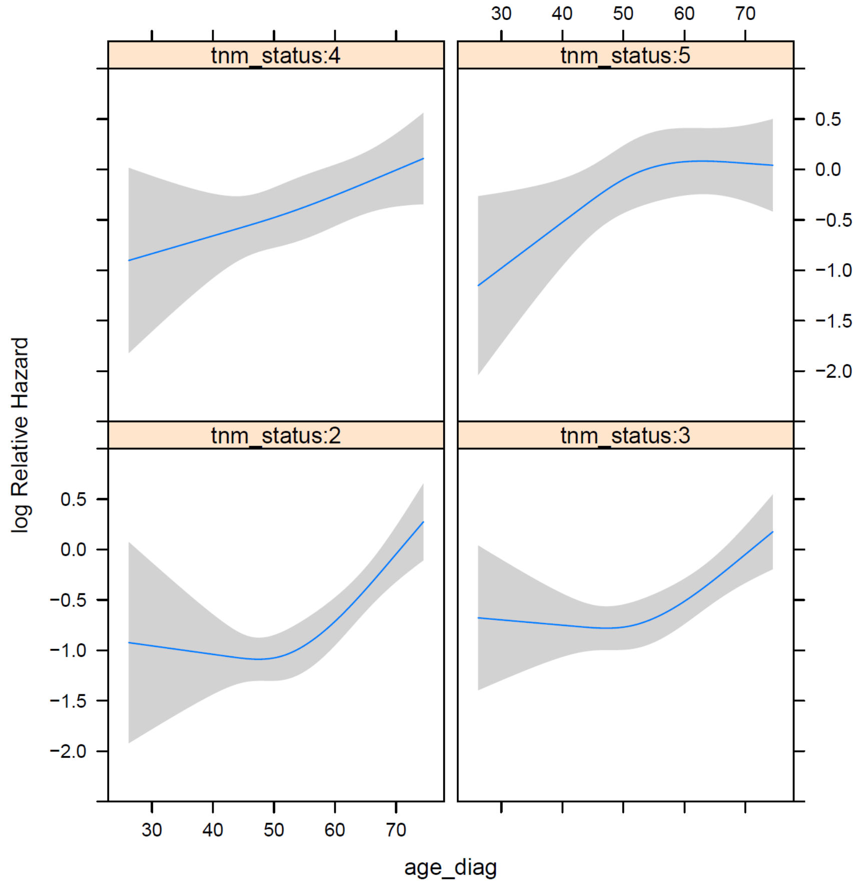


Figure 21: Effect of the Interaction of ER Status and Tamoxifen Therapy in 10-year Overall Survival Model

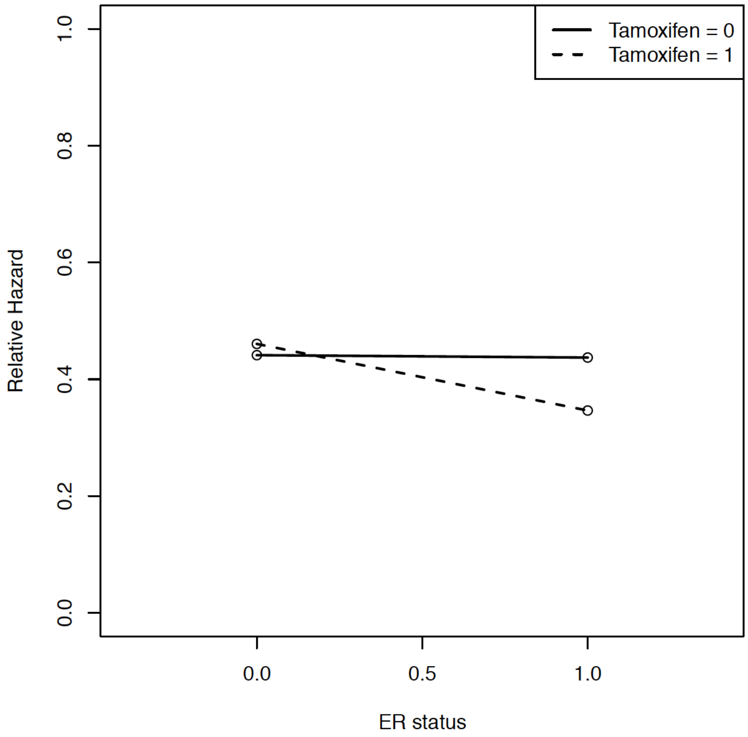


Figure 22: Effect of the Interaction of Age and Isoflavones in 10-year Overall Survival Model

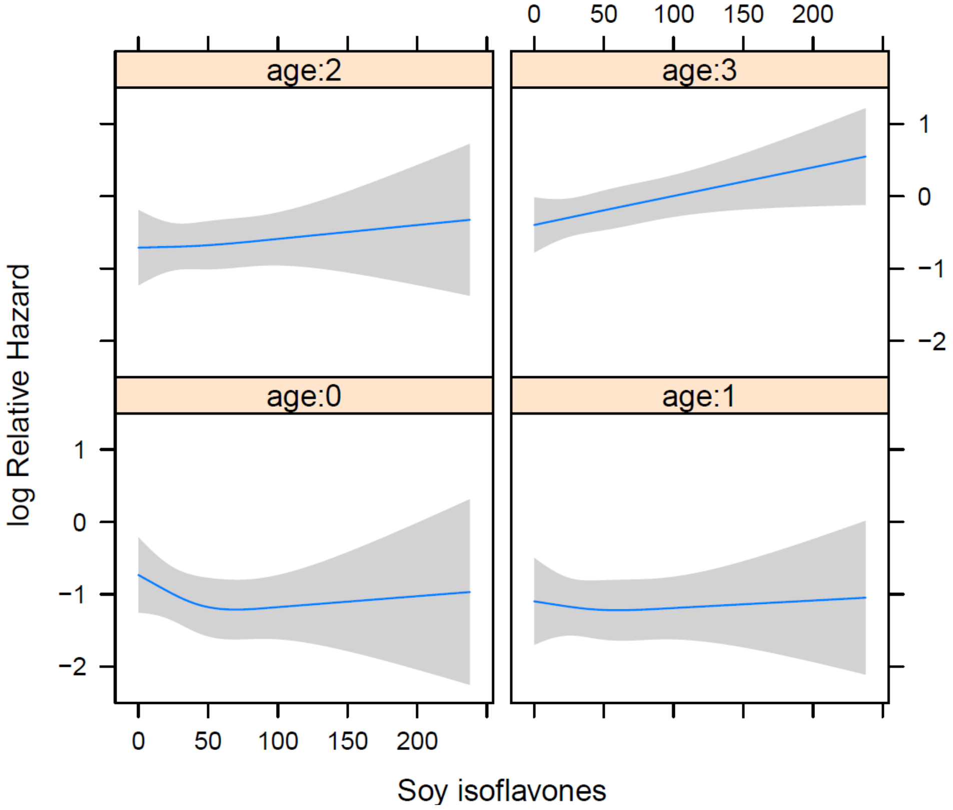


Figure 23: Survival Curve of 10-year Overall Survival Model

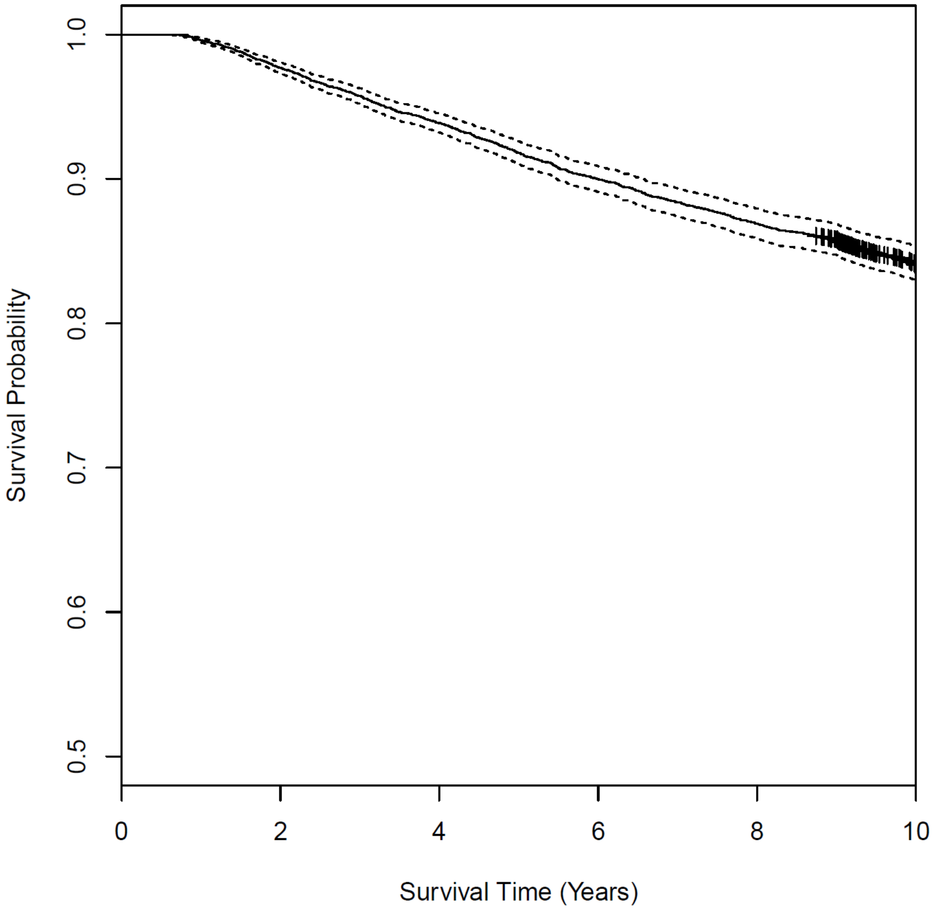


Figure 24: Calibration of 10-year Overall Survival Model

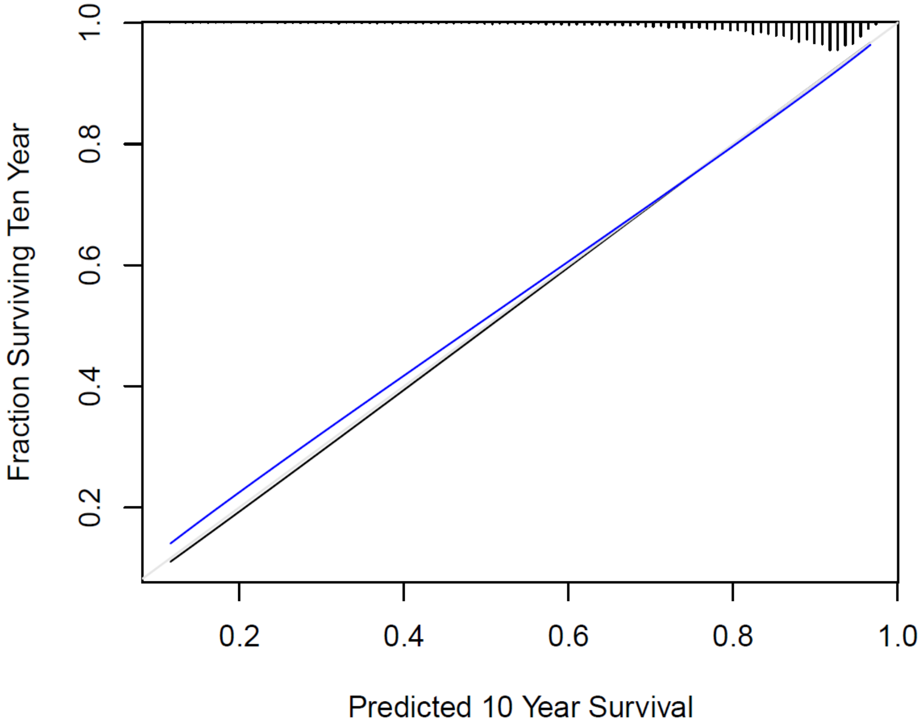
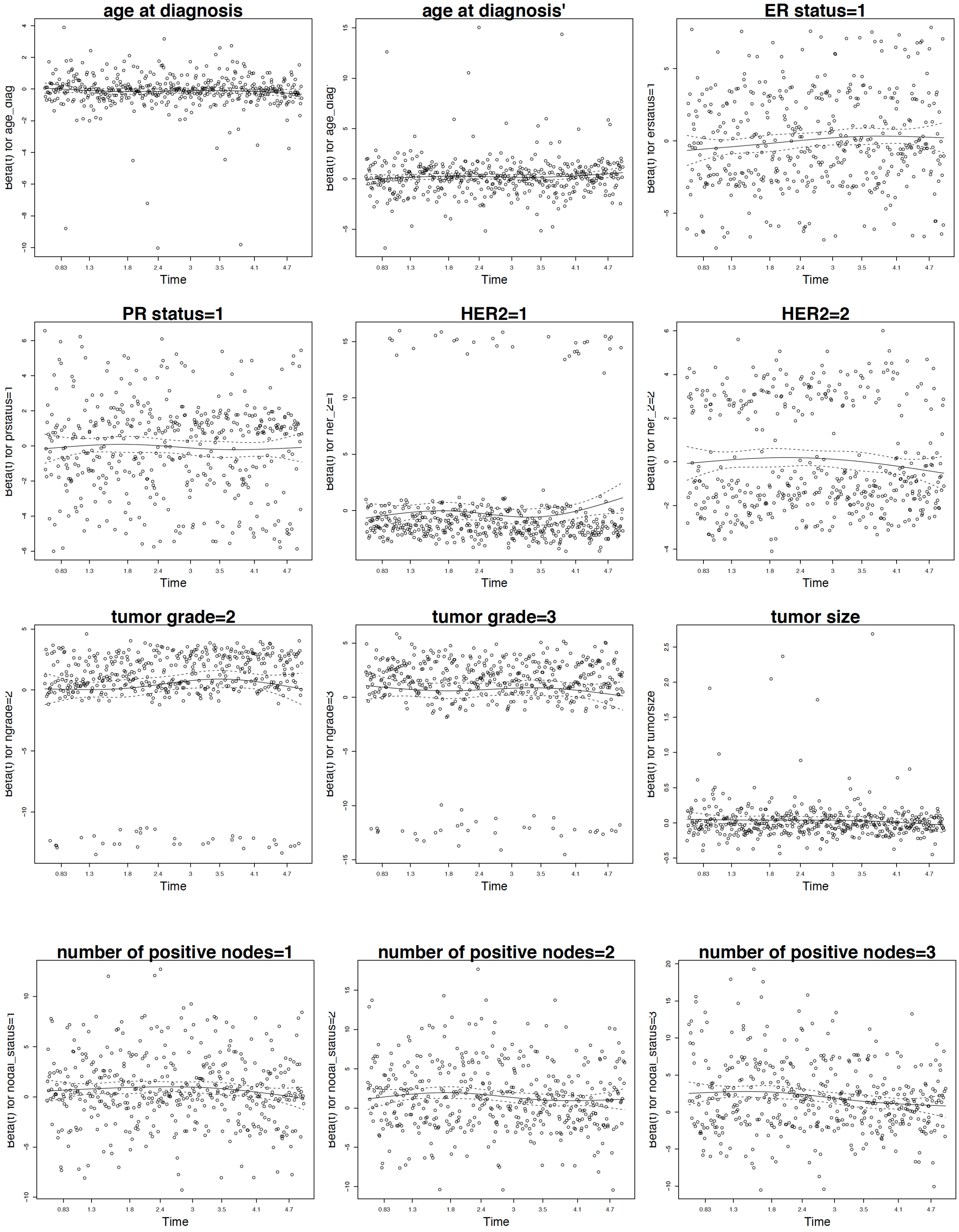
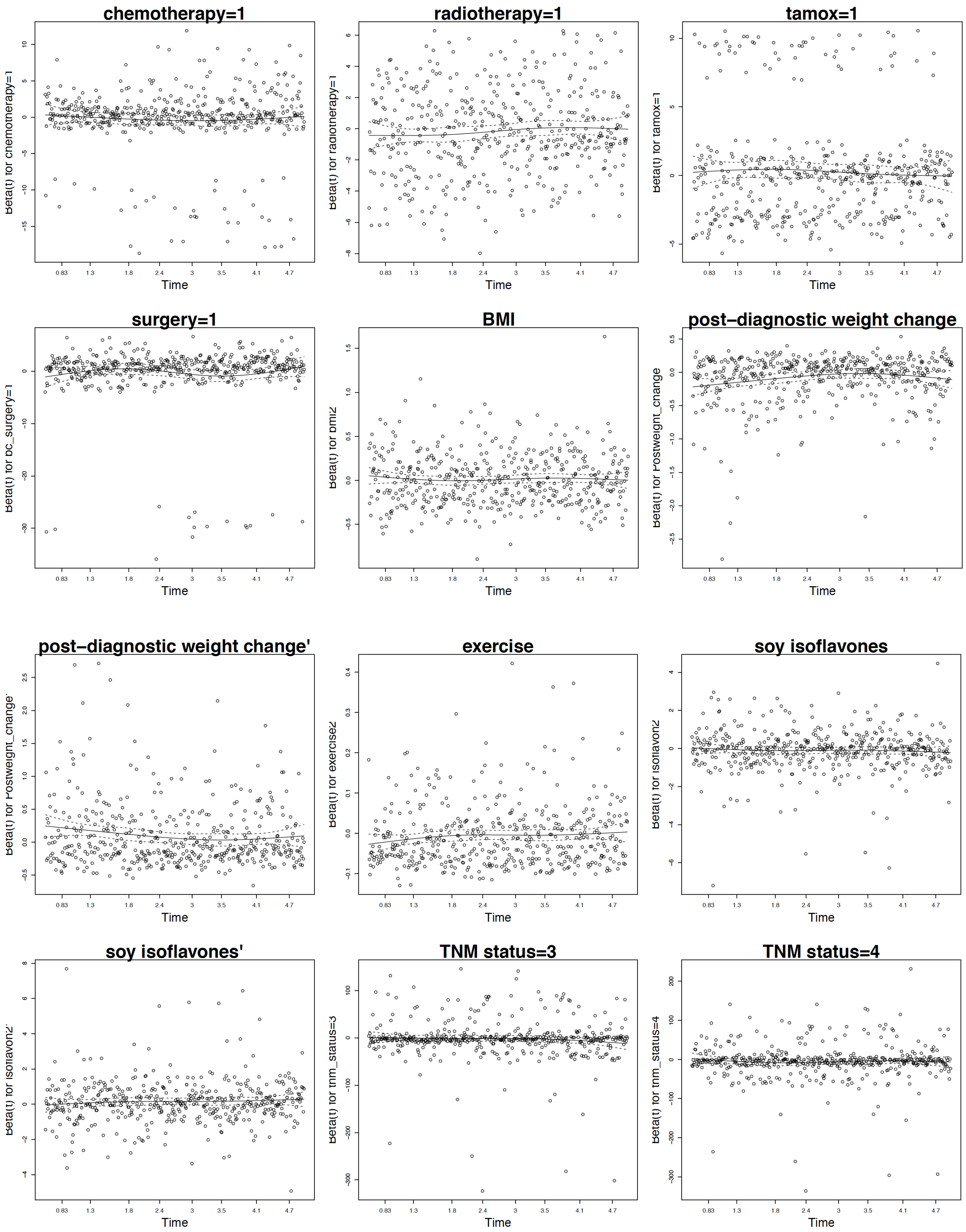


Figure 25: Factors Retained in Approximate Model of 10-year Overall Survival Model

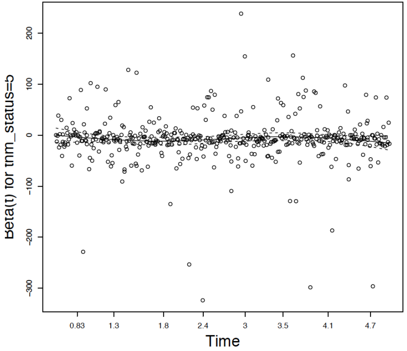
	Demographic Predictors			Pathological Predictors			Clinical Predictors				Treatment Predictors				Lifestyle Predictors		Interaction terms		
	age at diagnosis	BMI	Post-diagnostic weight change	ER status	PR status	HER2 status	Tumor grade	Tumor size	# of Nodes	Tumor-Node -Metastasis (TNM)	Chemotherapy	Radiotherapy	Tamoxifen	Surgery	Exercise	Isoflavones	age * TNM	ER * Tamoxifen	age * isoflavones
Imputation #1	●		●	●	●		●		●	●	●		●			●	●	●	●
Imputation #2	●		●	●	●		●	●	●	●	●		●		●		●	●	
Imputation #3	●		●	●			●	●	●	●	●	●	●		●		●	●	
Imputation #4	●		●	●	●		●		●	●	●		●		●	●	●	●	●
Imputation #5	●		●	●			●		●	●	●		●		●	●	●	●	●
Imputation #6	●		●	●	●		●	●	●	●	●		●		●	●	●	●	●
Imputation #7	●		●	●			●		●	●	●	●	●		●	●	●	●	●
Imputation #8	●		●	●			●		●	●	●	●	●		●	●	●	●	●
Imputation #9	●		●	●			●		●	●	●	●	●			●	●	●	●
Imputation #10	●		●	●			●		●	●	●	●	●		●	●	●	●	●
Count	10	0	10	10	4	0	10	3	10	10	10	5	10	0	8	8	10	10	8

Figure 26: Schoenfeld Residuals of Individual Predictors in 5-year Relapse-free Survival Model

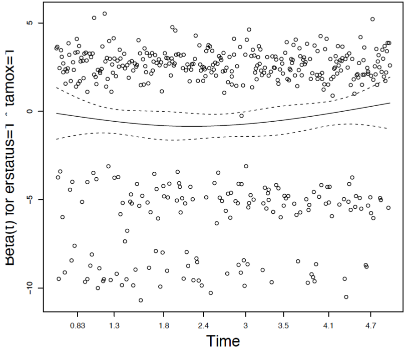



chemotherapy=1
Beta(t) for chemotherapy=1
10
5
0
-5
-10
-15
radiotherapy=1
Beta(t) for radiotherapy=1
6
4
2
0
-2
-4
-6
-8
tamox=1
Beta(t) for tamox=1
10
5
0
-5
surgery=1
Beta(t) for bc_surgery=1
0
-10
-20
-30
BMI
Beta(t) for bmi2
1.5
1.0
0.5
0.0
-0.5
post-diagnostic weight change
Beta(t) for Postweight_change
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
post-diagnostic weight change'
Beta(t) for Postweight_change'
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
exercise
Beta(t) for exercise2
0.4
0.3
0.2
0.1
0.0
-0.1
soy isoflavones
Beta(t) for isoflavon2
4
2
0
-2
-4
-6
soy isoflavones'
Beta(t) for isoflavon2'
8
6
4
2
0
-2
-4
TNM status=3
Beta(t) for tnm_status=3
100
0
-100
-200
-300
TNM status=4
Beta(t) for tnm_status=4
200
100
0
-100
-200
-300
0.83
1.3
1.8
2.4
3
3.5
4.1
4.7
Time

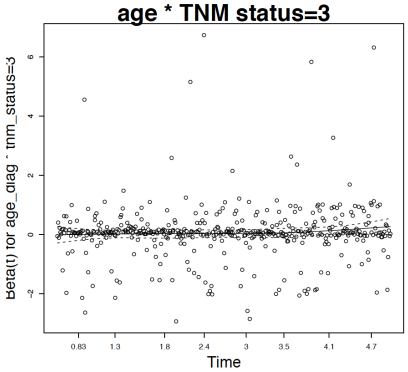
TNM status=5



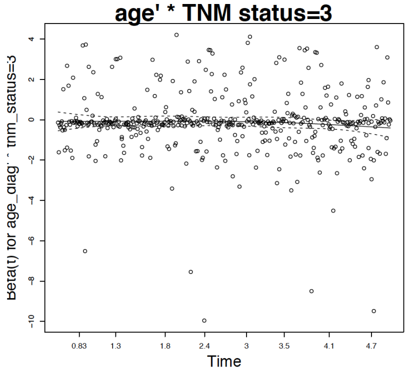
ER status=1 * tamoxifen=1



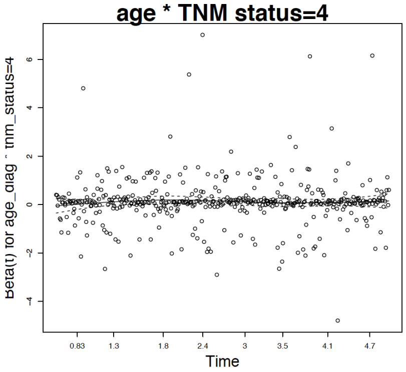
age * TNM status=3



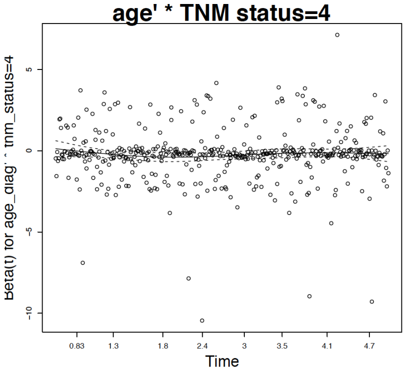
age' * TNM status=3



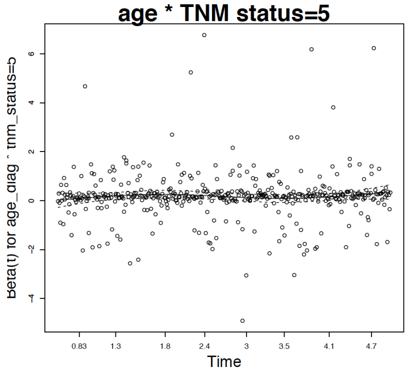
age * TNM status=4



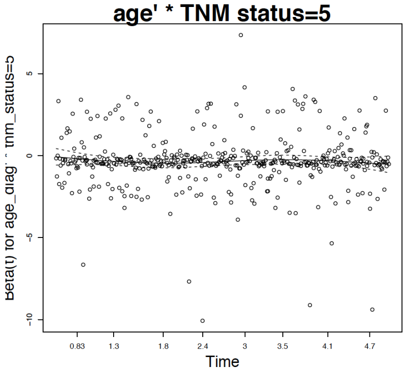
age' * TNM status=4



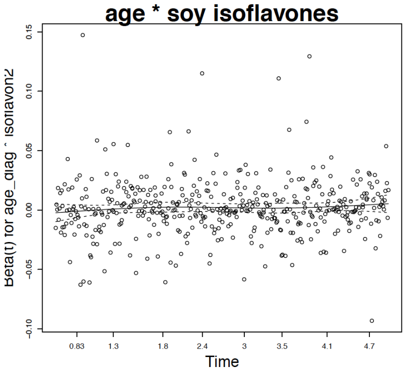
age * TNM status=5



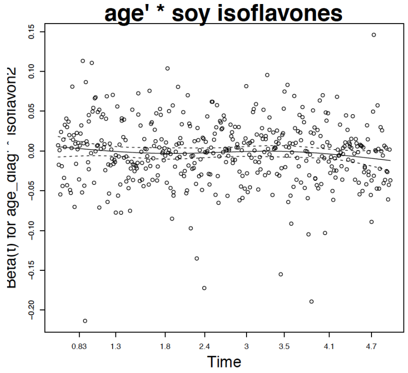
age' * TNM status=5



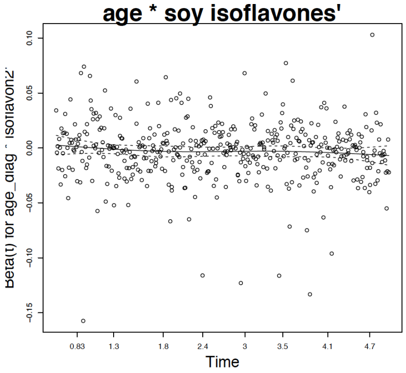
age * soy isoflavones



age' * soy isoflavones



age * soy isoflavones'



age' * soy isoflavones'

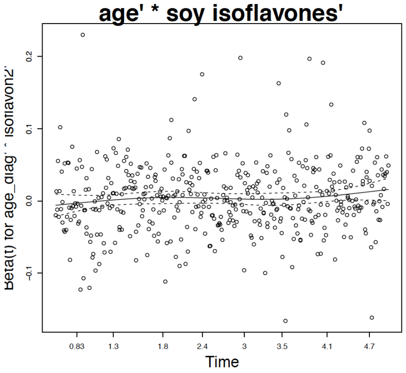


Figure 27. Nomogram of Predicting Median and Mean Survival Time in 5-year Relapse-free Survival Model

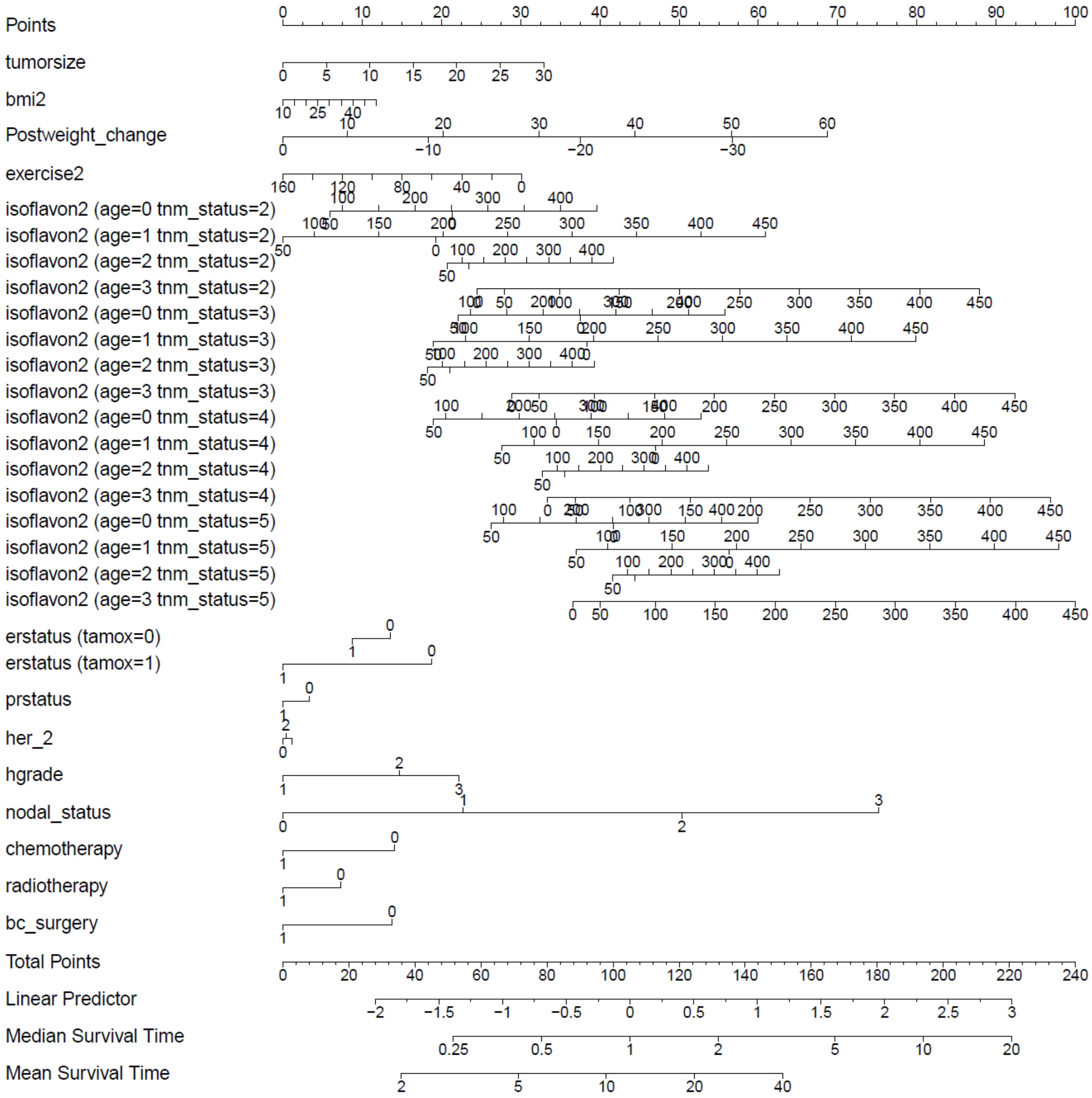


Figure 28: Contribution of Each Variable in 5-yr Relapse-free Survival Model

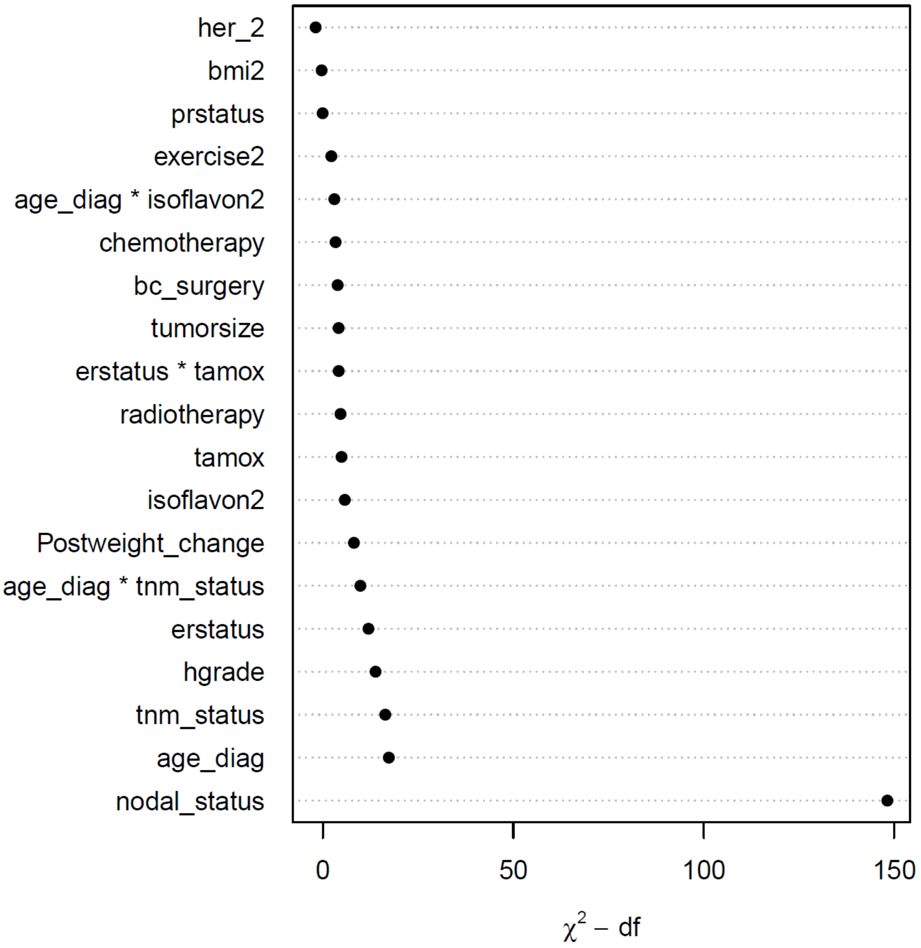


Figure 29: Estimated Hazard Ratios in 5-year Relapse-free Survival Model

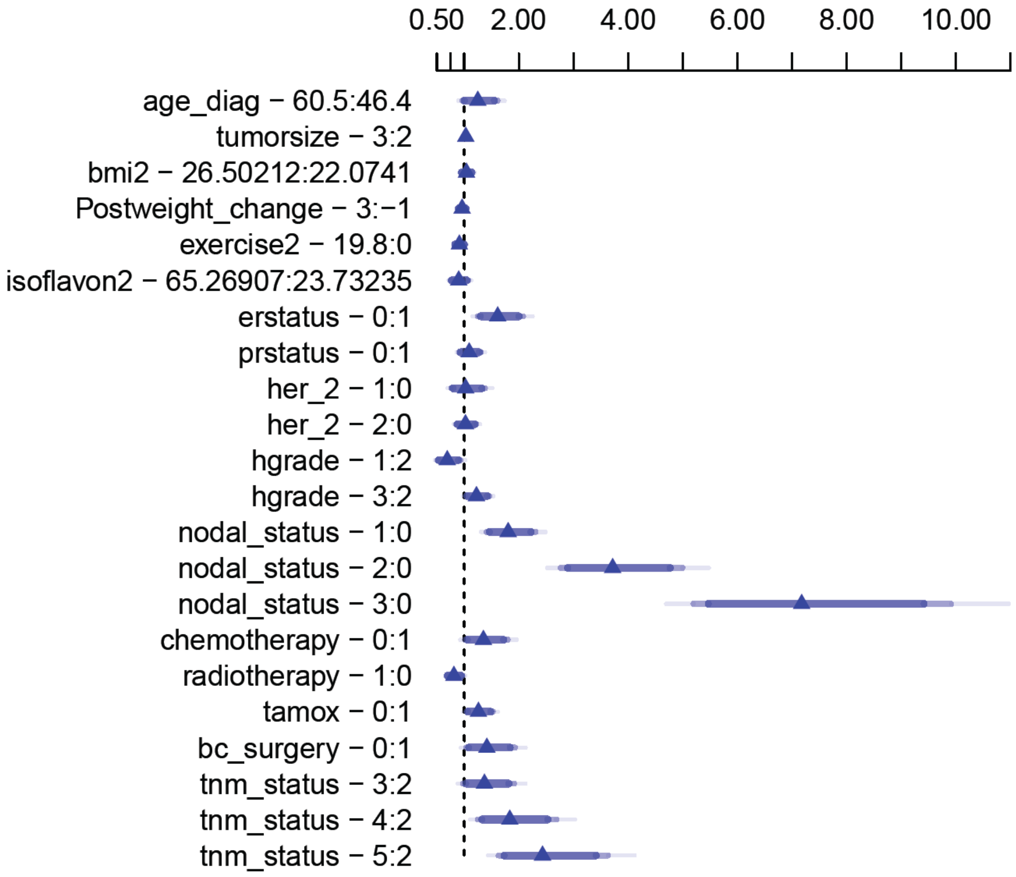


Figure 30: Effect of Each Predictor in 5-year Relapse-free Survival Model

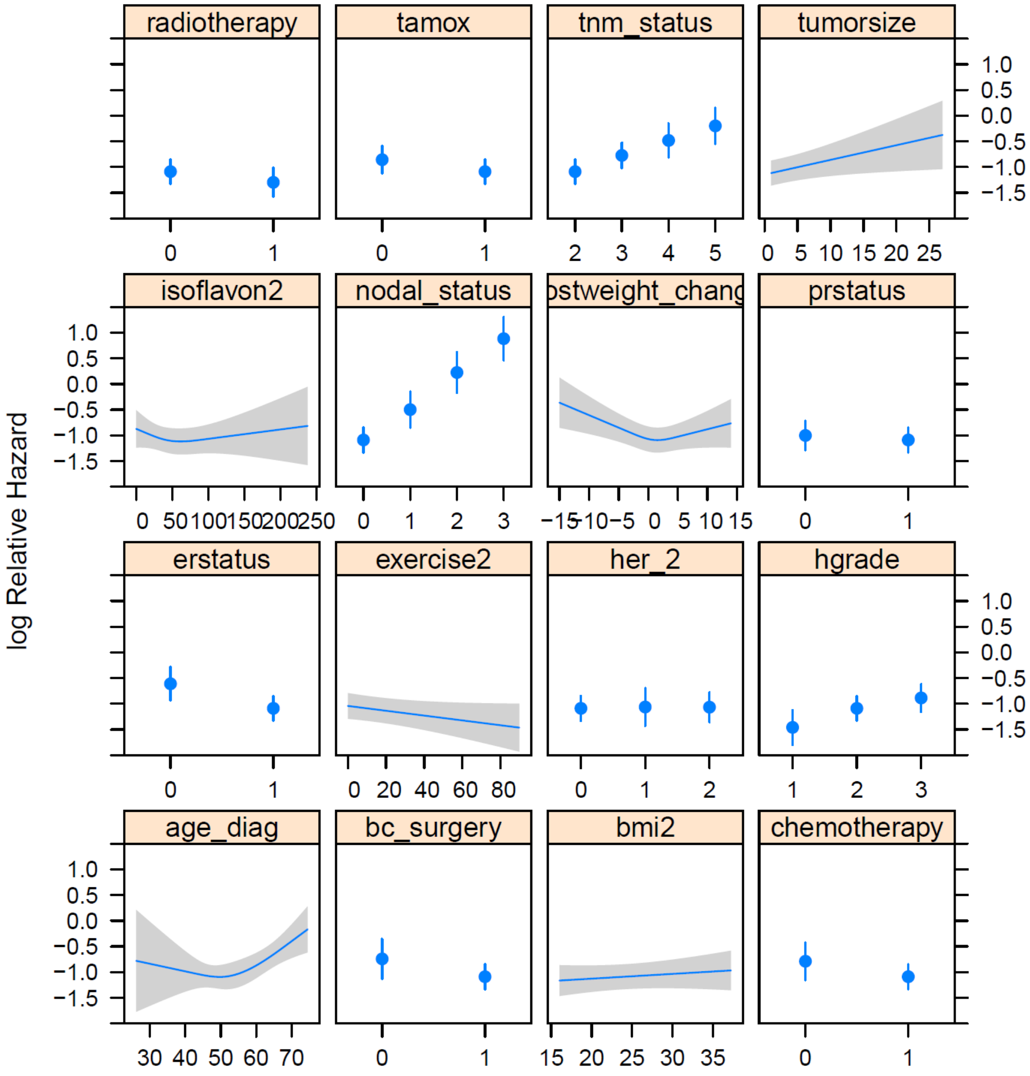


Figure 31: Effect of the Interaction of Age and Tumor-Node-Metastatic Status (TNM) in 5-year Relapse-free Survival Model

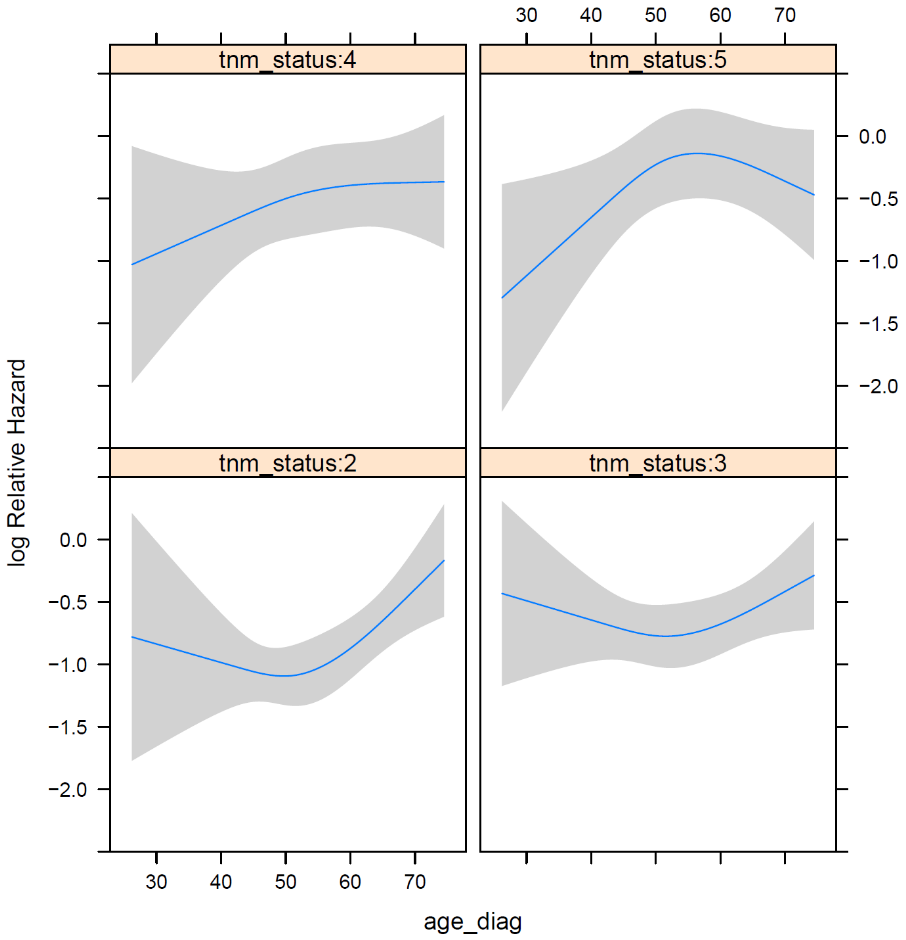


Figure 32: Effect of the Interaction of ER Status and Tamoxifen Therapy in 5-year Relapse-free Survival Model

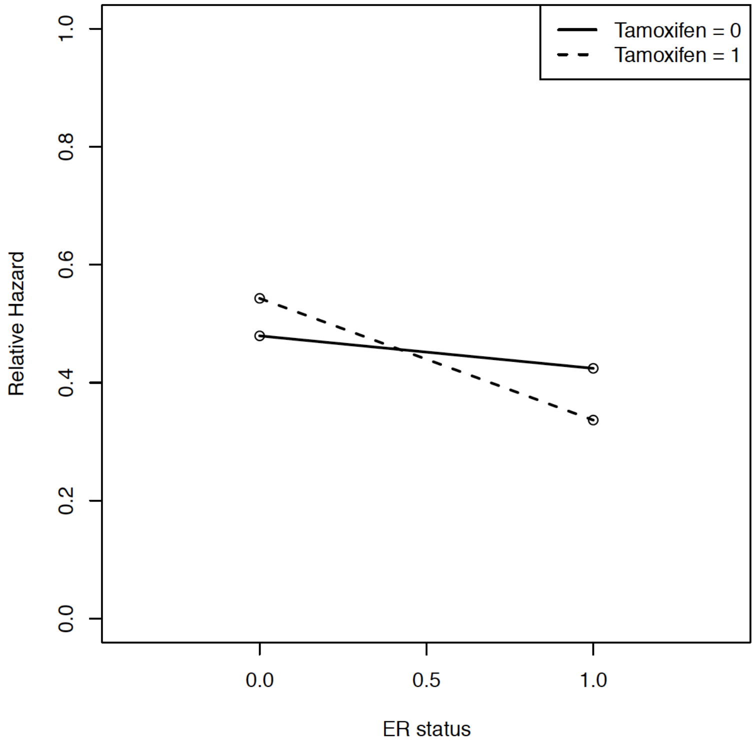


Figure 33: Effect of the Interaction of Age and Isoflavones in 5-year Relapse-free Survival Model

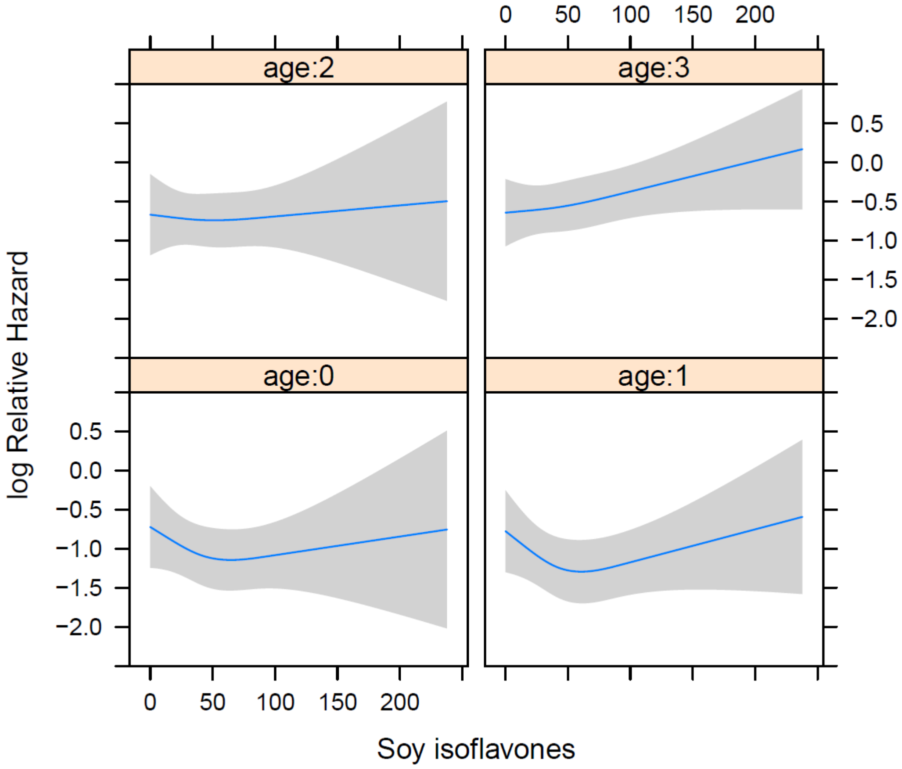


Figure 34: Survival Curve of 5-year Relapse-free Survival Model

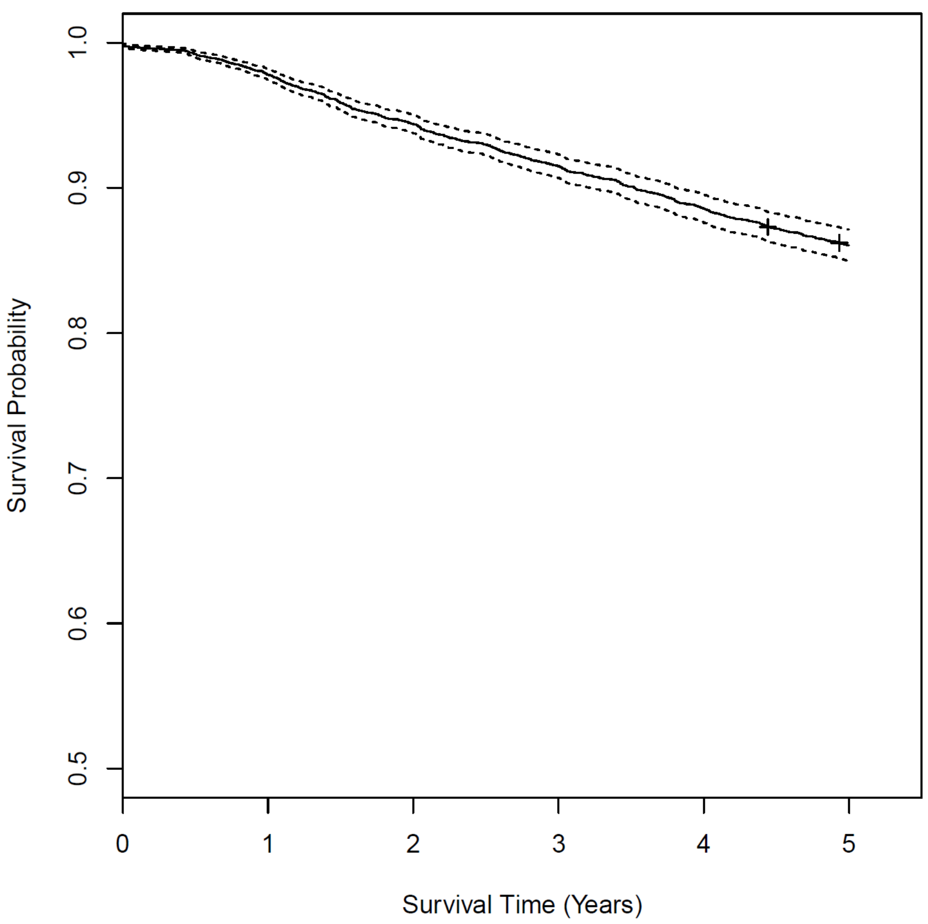


Figure 35: Calibration of 5-year Relapse-free Survival Model

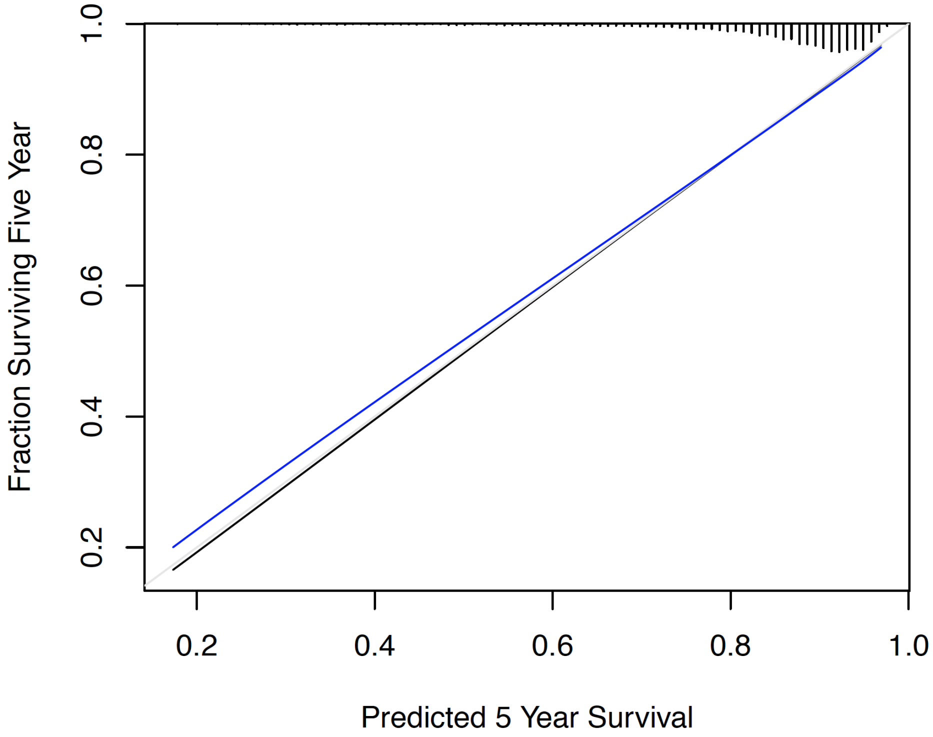
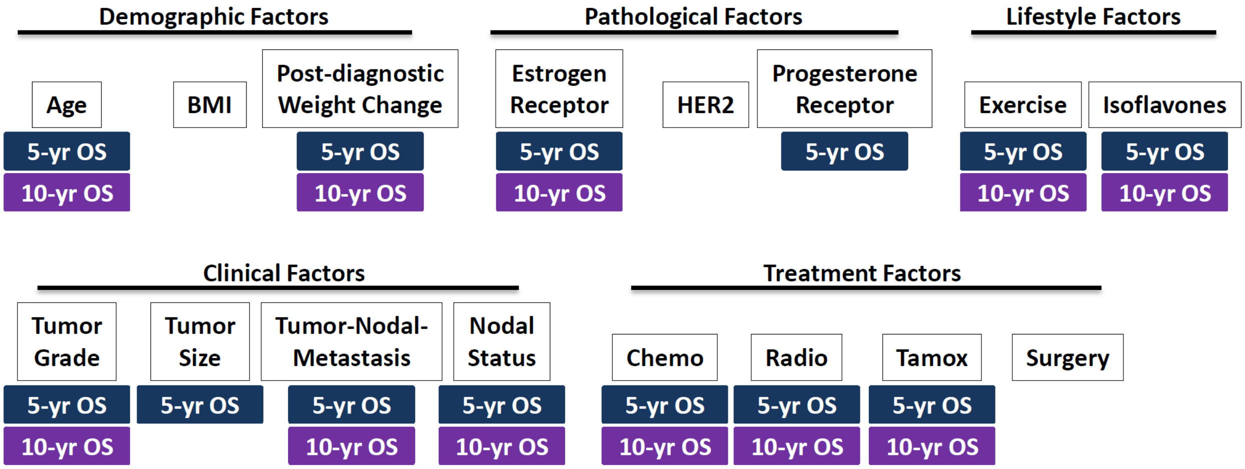


Figure 36: Factors Retained in Approximate Model of 5-year Relapse-free Survival Model

	Demographic Predictors			Pathological Predictors			Clinical Predictors				Treatment Predictors				Lifestyle Predictors		Interaction terms		
	age at diagnosis	BMI	Post-diagnostic weight change	ER status	PR status	HER2 status	Tumor grade	Tumor size	# of Nodes	Tumor-Node -Metastasis (TNM)	Chemotherapy	Radiotherapy	Tamoxifen	Surgery	Exercise	Isoflavones	age * TNM	ER * Tamoxifen	age * isoflavones
Imputation #1	●		●	●			●		●	●	●	●	●	●	●	●	●	●	●
Imputation #2	●		●	●			●		●	●	●	●	●	●	●	●	●	●	●
Imputation #3	●		●	●			●		●	●	●	●	●	●	●	●	●	●	●
Imputation #4	●		●	●			●	●	●	●	●	●	●	●	●	●	●	●	●
Imputation #5	●		●	●			●	●	●	●	●	●	●	●		●	●	●	●
Imputation #6	●		●	●			●		●	●	●	●	●	●	●	●	●	●	●
Imputation #7	●		●	●			●	●	●	●	●	●	●	●	●	●	●	●	●
Imputation #8	●		●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●
Imputation #9	●		●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●
Imputation #10	●		●	●			●		●	●	●	●	●	●	●	●	●	●	●
Count	10	0	10	10	2	0	10	5	10	10	10	10	10	10	9	10	10	10	10

Figure 37: Comparison of Approximate Models for 5-year and 10-year Overall Survival



Appendix:

Figure S1: Mathematical Form to Estimate Log Hazard of 5-year Overall Survival –Full Model

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

$$
\begin{array}{l}
X\hat{\beta} = \\
\quad 0.7711398 \\
\quad -0.02457168\mathrm{age}_{\mathrm{d}}\mathrm{iag} + 7.377405\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_{+}^{3} \\
\quad -0.0001082554(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_{+}^{3} + 3.448135\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_{+}^{3} \\
\quad -0.1276254[1] \\
\quad -0.2543161[1] \\
\quad +0.03080469[1] + 0.02484729[2] \\
\quad +0.6733415[2] + 0.9596952[3] + 0.0336742\ \mathrm{tumorsize} \\
\quad +0.6433588[1] + 1.41801[2] + 2.14201[3] \\
\quad -0.3678594[1] \\
\quad -0.1948056[1] \\
\quad +0.2746814[1] \\
\quad -0.2033122[1] + 0.002105633\ \mathrm{bmi}_2 \\
\quad -0.07114351\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} + 0.001147811(\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} + 3)_{+}^{3} \\
\quad -0.002295623(\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} - 1)_{+}^{3} \\
\quad +0.001147811(\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} - 5)_{+}^{3} - 0.00370681\ \mathrm{exercise}_2 \\
\quad -0.008900646\mathrm{isoflavon}_2 - 3.394975\times10^{-6}(\mathrm{isoflavon}_2 - 11.57483)_{+}^{3} \\
\quad +5.358482\times10^{-6}(\mathrm{isoflavon}_2 - 41.27154)_{+}^{3} \\
\quad -1.963507\times10^{-6}(\mathrm{isoflavon}_2 - 92.61826)_{+}^{3} \\
\quad +0.7953047[3] - 0.9064132[4] - 1.493399[5] \\
\quad -0.6272337\ [1]\ \times\ [1] \\
\quad +[3][-0.01258545\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 3.527301\times10^{-6}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_{+}^{3} \\
\quad +5.175931\times10^{-6}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_{+}^{3} - 1.64863\times10^{-6}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_{+}^{3}] \\
\quad +[4][0.03049839\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 0.0001321936(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_{+}^{3} \\
\quad +0.0001939797(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_{+}^{3} - 6.178614\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_{+}^{3}] \\
\quad +[5][0.04704\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 0.0001694673(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_{+}^{3} \\
\quad +0.0002486749(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_{+}^{3} - 7.920755\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_{+}^{3}] \\
\quad +\mathrm{age}_{\mathrm{d}}\mathrm{iag}[3.862235\times10^{-5}\ \mathrm{isoflavon}_2 + 8.720164\times10^{-8}(\mathrm{isoflavon}_2 - 11.57483)_{+}^{3} \\
\quad -1.376353\times10^{-7}(\mathrm{isoflavon}_2 - 41.27154)_{+}^{3} \\
\quad +5.043366\times10^{-8}(\mathrm{isoflavon}_2 - 92.61826)_{+}^{3}] \\
\quad +\mathrm{age}_{\mathrm{d}}\mathrm{iag}'[0.000901362\mathrm{isoflavon}_2 \\
\quad -2.088448\times10^{-7}(\mathrm{isoflavon}_2 - 11.57483)_{+}^{3} \\
\quad +3.296316\times10^{-7}(\mathrm{isoflavon}_2 - 41.27154)_{+}^{3} \\
\quad -1.207868\times10^{-7}(\mathrm{isoflavon}_2 - 92.61826)_{+}^{3}]
\end{array}
$$

Figure S2: Mathematical Form to Estimate Log Hazard of 10-year Overall Survival Model

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

$$\begin{aligned}
X\hat{\beta} = & \\
& 2.22578 \\
& -0.06072926\mathrm{age}_{\mathrm{d}}\mathrm{iag} + 0.0001951375(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_+^3 \\
& -0.0002863431(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_+^3 + 9.120557\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_+^3 \\
& -0.00923343[1] \\
& -0.1328945[1] \\
& -0.05933487[1] + 0.01726178[2] \\
& +0.4428396[2] + 0.6283434[3] + 0.02519524\ \mathrm{tumorsize} \\
& +0.5627348[1] + 1.200651[2] + 1.837779[3] \\
& -0.3943456[1] \\
& -0.1753997[1] \\
& +0.04272414[1] \\
& -0.119265[1] + 0.005933787\ \mathrm{bmi}_2 \\
& -0.05439115\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} + 0.000870575(\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} + 3)_+^3 \\
& -0.00174115(\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} - 1)_+^3 \\
& +0.000870575(\mathrm{Postweight}_{\mathrm{c}}\mathrm{hange} - 5)_+^3 - 0.005849782\ \mathrm{exercise}_2 \\
& -0.07082239\mathrm{isoflavon}_2 + 1.18304\times10^{-5}(\mathrm{isoflavon}_2 - 11.57483)_+^3 \\
& -1.867259\times10^{-5}(\mathrm{isoflavon}_2 - 41.27154)_+^3 \\
& +6.84219\times10^{-6}(\mathrm{isoflavon}_2 - 92.61826)_+^3 \\
& +0.1629577[3] - 0.6621874[4] - 1.641744[5] \\
& -0.2753672\ [1]\ \times\ [1] \\
& +[3][0.003150101\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 3.541945\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_+^3 \\
& +5.19742\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_+^3 - 1.655475\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_+^3] \\
& +[4][0.02606242\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 0.000103127(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_+^3 \\
& +0.0001513277(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_+^3 - 4.820068\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_+^3] \\
& +[5][0.05392161\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 0.000186041(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 42.5)_+^3 \\
& +0.000272995(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 51.1)_+^3 - 8.695396\times10^{-5}(\mathrm{age}_{\mathrm{d}}\mathrm{iag} - 69.5)_+^3] \\
& +\mathrm{age}_{\mathrm{d}}\mathrm{iag}[0.00141388\mathrm{isoflavon}_2 - 2.352472\times10^{-7}(\mathrm{isoflavon}_2 - 11.57483)_+^3 \\
& +3.71304\times10^{-7}(\mathrm{isoflavon}_2 - 41.27154)_+^3 \\
& -1.360568\times10^{-7}(\mathrm{isoflavon}_2 - 92.61826)_+^3] \\
& +\mathrm{age}_{\mathrm{d}}\mathrm{iag}'[-0.001622516\mathrm{isoflavon}_2 \\
& +3.151504\times10^{-7}(\mathrm{isoflavon}_2 - 11.57483)_+^3 \\
& -4.974197\times10^{-7}(\mathrm{isoflavon}_2 - 41.27154)_+^3 \\
& +1.822694\times10^{-7}(\mathrm{isoflavon}_2 - 92.61826)_+^3]
\end{aligned}$$

Figure S3: Mathematical Form to Estimate Log Hazard of 5-year Relapse-free Survival Model

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

$$
\begin{array}{l}
X\hat{\beta} = \\
\quad 1.809649 \\
\quad -0.04453366\text{age}_{\text{d}}\text{iag} + 0.0001182779(\text{age}_{\text{d}}\text{iag} - 42.5)_{+}^{3} \\
\quad -0.00017356(\text{age}_{\text{d}}\text{iag} - 51.1)_{+}^{3} + 5.528208\times10^{-5}(\text{age}_{\text{d}}\text{iag} - 69.5)_{+}^{3} \\
\quad -0.122059[1] \\
\quad -0.08728378[1] \\
\quad +0.02572236[1] + 0.02276423[2] \\
\quad +0.3710397[2] + 0.5734821[3] + 0.02854179 \text{ tumorsize} \\
\quad +0.5903483[1] + 1.313109[2] + 1.970961[3] \\
\quad -0.3010138[1] \\
\quad -0.2088536[1] \\
\quad +0.1245148[1] \\
\quad -0.3469829[1] + 0.009261416 \text{ bmi}_2 \\
\quad -0.04855723\text{Postweight}_{\text{c}}\text{hange} + 0.000807353(\text{Postweight}_{\text{c}}\text{hange} + 3)_{+}^{3} \\
\quad -0.001614706(\text{Postweight}_{\text{c}}\text{hange} - 1)_{+}^{3} \\
\quad +0.000807353(\text{Postweight}_{\text{c}}\text{hange} - 5)_{+}^{3} - 0.004695954 \text{ exercise}_2 \\
\quad -0.04707053\text{isoflavon}_2 + 8.762083\times10^{-6}(\text{isoflavon}_2 - 11.67096)_{+}^{3} \\
\quad -1.373902\times10^{-5}(\text{isoflavon}_2 - 41.26368)_{+}^{3} \\
\quad +4.976935\times10^{-6}(\text{isoflavon}_2 - 93.36277)_{+}^{3} \\
\quad +0.3621417[3] - 1.233378[4] - 2.124969[5] \\
\quad -0.3561068\ [1]\ \times\ [1] \\
\quad +[3][-0.0005272039\text{age}_{\text{d}}\text{iag} - 3.151915\times10^{-5}(\text{age}_{\text{d}}\text{iag} - 42.5)_{+}^{3} \\
\quad +4.625093\times10^{-5}(\text{age}_{\text{d}}\text{iag} - 51.1)_{+}^{3} - 1.473178\times10^{-5}(\text{age}_{\text{d}}\text{iag} - 69.5)_{+}^{3}] \\
\quad +[4][0.03757299\text{age}_{\text{d}}\text{iag} - 0.0001258324(\text{age}_{\text{d}}\text{iag} - 42.5)_{+}^{3} \\
\quad +0.0001846454(\text{age}_{\text{d}}\text{iag} - 51.1)_{+}^{3} - 5.881299\times10^{-5}(\text{age}_{\text{d}}\text{iag} - 69.5)_{+}^{3}] \\
\quad +[5][0.06143818\text{age}_{\text{d}}\text{iag} - 0.0001964681(\text{age}_{\text{d}}\text{iag} - 42.5)_{+}^{3} \\
\quad +0.0002882955(\text{age}_{\text{d}}\text{iag} - 51.1)_{+}^{3} - 9.182746\times10^{-5}(\text{age}_{\text{d}}\text{iag} - 69.5)_{+}^{3}] \\
\quad +\text{age}_{\text{d}}\text{iag}[0.0008142998\text{isoflavon}_2 \\
\quad -1.532109\times10^{-7}(\text{isoflavon}_2 - 11.67096)_{+}^{3} \\
\quad +2.40236\times10^{-7}(\text{isoflavon}_2 - 41.26368)_{+}^{3} \\
\quad -8.702506\times10^{-8}(\text{isoflavon}_2 - 93.36277)_{+}^{3}] \\
\quad +\text{age}_{\text{d}}\text{iag}'[-0.0005103616\text{isoflavon}_2 \\
\quad +1.518659\times10^{-7}(\text{isoflavon}_2 - 11.67096)_{+}^{3} \\
\quad -2.38127\times10^{-7}(\text{isoflavon}_2 - 41.26368)_{+}^{3} \\
\quad +8.62611\times10^{-8}(\text{isoflavon}_2 - 93.36277)_{+}^{3}]
\end{array}
$$