Final Case Study - Breast Cancer Survival Analysis Under Different Treatments

Piotr M. Suder

Abstract

Breast Invasive Ductal Carcinoma (BIDC) is the most prevalent type of breast cancer, comprising over 70% of all cases. The two primary surgical treatments include mastectomy and breast-conserving surgery (BCS). In this study we perform statistical inference on the observational data on patient survival provided by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), in order to compare the effectiveness of these two procedures in treating BIDC. In order to account for various confounders such as patient and tumor characteristics, other treatment components, and patients' genomes, we employ the Cox proportional hazards model for our analysis. Following Mao et al, we estimate the restricted average survival causal effect (RACE). Our results suggest that BCS might result in reduced hazard compared to mastectomy, however we do not have conclusive statistical evidence for that. Additionally, we identify several genes whose expression levels have statistically significant contribution to heterogeneity in treatment outcomes between patients.

1 Introduction

Breast cancer is the most common cancer among women worldwide, accounting for 30% of all cancer cases in this group and causing significant mortality. It is therefore crucial to optimize treatment strategies that can improve patient outcomes, reduce complications, and enhance quality of life. Two primary types of surgeries performed in order to remove breast cancer are mastectomy and breast-conserving surgery (BCS). In many cases, BCS might be preferred, especially in earlier stages of cancer, since it avoids removing the entire breast. However, the decision-making process is complex and involves a careful assessment of the potential risks and benefits of each surgical option.

In this context, the effectiveness of both procedures in prolonging life expectation and reducing risk of death is a crucial consideration in the choice of treatment. In this report we analyze the survival times in the observational study of a group of patients who underwent mastectomy or BCS, while accounting for confounders such as other components of the treatment, patient characteristics, tumor characteristics, and patients' genomes. We focus on patients suffering from breast invasive ductal carcinoma (BIDC), the most common type of breat cancer accounting for more than 70% of cases. We consider the individuals in stages 1,2 and 3 of tumor development, since BCS is very rarely recommended in treatment of late stage cancer and there are only two patients with stage 0 in the dataset.

From the point of view of causal inference we are interested in estimating the total **restricted average survival causal effect (RACE)** which Mao et al define as

$$\Delta_{\text{RACE}}(t^*) = \int_0^{t^*} S_1(t)dt - \int_0^{t^*} S_0(t)dt \tag{1}$$

where S_Z is the population survival function for the case when the entire population is given treatment Z. In this case we take t^* to be the longest observed survival or censoring time, whichever is longer. We are also interested in identifying sources of potential heterogeneity in treatment outcomes between groups of patients based on their characteristics and other treatment components involved.

2 Exploratory Data Analysis

We perform the analysis on the dataset provided by Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). We select a subset of patients with more common genetic mutations for BIDC (those mutations which occur in more than 1.3% of the population in the dataset). While this cutoff might seem somewhat arbitrary it serves to balance the inclusion of the largest possible number of mutations in the study with trying to avoid having very rare classes of patients which could lead to problems with cross-validation and instability in the model parameter estimation later on. This leaves us with 889 patients, among whom 522 had mastectomy and 367 underwent the BCS procedure. For each of them we have the survival time recorded in months and the censoring status. We also have information about the patient characteristics such as age, menopause status, number of lymph nodes determined to be positive for cancer, Nottingham Prognostic Index, other types of treatment involved such as chemotherapy, hormone and radio therapy, tumor characteristics such as stage, size, molecular subtype, HER2 status, ER status, cellularity, neoplasm histologic subtype, integrative cluster, primary laterality, gene classifier subtype, as well as pam 50 amd claudin-low subtype. We also have expression levels for 489 genes and mutation indicators for 173 of those genes. This gives a total of 681 covariates, which after introducing interactions with the treatment, leads to a very high dimensional survival modeling problem. As shown in Figure 1, we observed death of 39.8% of the patients who had mastectomy and 27.5% of those who undergone BCS. The remaining individuals had their survival times censored. The right panel of Figure 1 shows the population survival functions depending on the treatment estimated according to the Kaplan-Meier method. As we can see, patients who received BCS tend to have higher estimated survival probability throughout most of the time. However, it is likely that there are some confounding variables which influence that result.

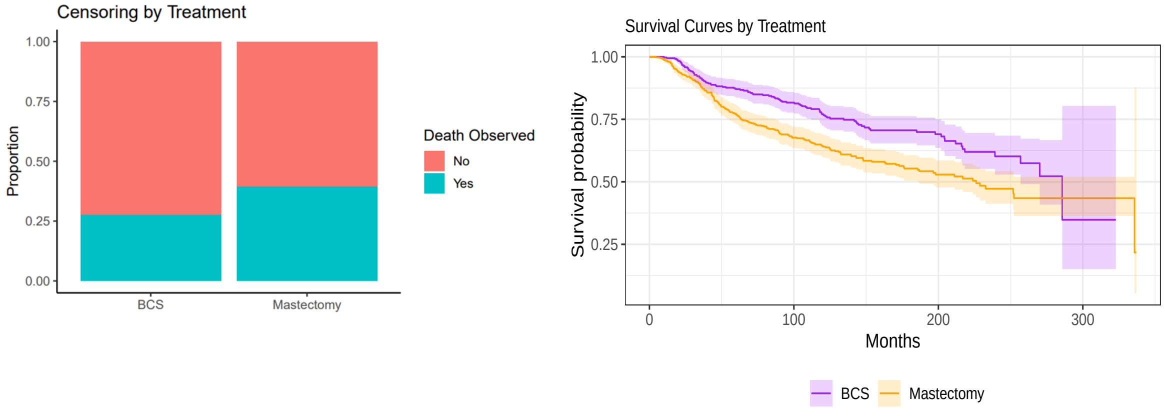

Figure 1: Censoring proportion for both treatment groups (left); estimated survival functions for mastectomy and breast conserving surgery (right).

A significant proportion (27.4%) of the observations have missing values. The majority of the missing values are for the tumor stage, which was not recorded for 25.4% of patients. Other covariates with missing values are neoplasm histologic grade (3.1% of observations) and tumor size (0.6% of observations). The fact that a large portion of tumor stage observations is missing poses some challenges, since we will only include patients with stages 1, 2 and 3 of tumor development in the analysis. Therefore, care must be taken when imputing the missing values, even though there are fewer than 1% of observations with tumor stages 0 or 4, so it is unlikely that a large number of individuals would have the tumor stage observation imputed to one of these values. We will run sensitivity analysis for data imputation in order to alleviate this issue. We assume the data to be missing at random (MAR) and use multivariate imputation by chained equations (MICE), following van Buuren and Groothuis-Oudshoorn for the imputation. We then exclude the observations with tumor stages 0 and 4, which leaves us with 879 individuals. We repeat the imputation five times for the purposes of sensitivity analysis which we will run at a later stage of the analysis. Figure 3 in the Appendix shows the population survival curves depending on the treatment estimated according to the Kaplan-Meier method for the three stages of tumor development we are considering. Interestingly, it seems that while for patients with stage 1 tumors the estimated survival probability over time seems to be almost exactly the same for both treatments, we see a significant difference for stage 2, where BCS is associated with higher probability of survival throughout most of the time. This is surprising since BCS in most cases is recommended in early stages of the tumor development. This could suggest that the comparative benefits of BCS might be larger for stage 2 than for stage 1 of tumor development. The results for stage 3 patients are more difficult to assess, since we have much fever patients who underwent the BCS in that group and thus there is large uncertainty in the survival function estimate.

3 Modeling and Analysis

Since we have a large number of categorical variables in the dataset we use the method proposed by Gelman to standardize the continous covariates by subtracting their means and dividing each of them by two times its standard deviation. We include the information about the gene mutations in the model by introducing interaction terms between the mutation indicators and the expression levels of the corresponding genes in addition to the main effect terms of the expression levels. We do not include the main effects terms of the mutation indicators. This way the strength of influence of a given mutation is proportional to how strongly the corresponding gene is expressed. In order to account for potential heterogeneity in the survival outcomes between the treatments caused by patient's characteristics and other factors we also introduce interaction terms between the treatment and other covariates.

3.1 Candidate Models

We considered two models for our analysis, the **Cox proportional hazards model** and the **semiparametric accelerated failure time (AFT) model**. Under the Cox model, we assume the hazard for the i-th individual at time t to be given by

$$h(t) = h_0(t)\exp(\boldsymbol{w}_i^T\boldsymbol{\beta}) \tag{2}$$

where $\boldsymbol{w}_i = [z_i, \boldsymbol{x}_i^T, z_i \cdot \boldsymbol{x}_i^T]^T$ and $\boldsymbol{\beta} \in \mathbb{R}^{1363}$, with $h_0(t)$ being the baseline hazard function, $\boldsymbol{x}_i$ being the vector of covariates for the i-th patient, and z_i being the indicator of the treatment (we set $z_i = 1$ for mastectomy and $z_i = 0$ for BCS).

Under the semiparametric accelerated failure time (AFT) model, we assume that the survival time T_i of the i-th individual is given by

$$\log(T_i) = \boldsymbol{w}_i^T\boldsymbol{\beta} + \epsilon_i \tag{3}$$

where $\boldsymbol{w}_i = [z_i, \boldsymbol{x}_i^T, z_i \cdot \boldsymbol{x}_i^T]^T$, $\boldsymbol{\beta} \in \mathbb{R}^{1363}$, and ϵ_i are i.i.d. mean-zero random variables. This is a semiparametric model since no specified distribution is assumed for the error terms ϵ_i.

3.2 Model Comparison

We split the dataset into training and testing data according to the 70/30 proportion. In order to deal with high dimensionality of the data we fit both models using the elastic net penalty with $\alpha = 0.5$, which puts equal weight on ridge and lasso penalties. We tune it using 5-fold cross-validation. We fit the Cox model using the method proposed by Simon et al with help of the `glmnet` package. The semiparametric AFT model is fitted via the penalized Gehan estimator following the method proposed by Suder and Molstad by using the `penAFT` package. Cox model gives out-of-sample concordance of 0.677, while semiparametric AFT has the concordance of 0.664. Both models have very similar performance, however, we choose to proceed with Cox, primarily for the computational efficiency reasons.

3.3 Propensity Scores and Variable Selection

We model the propensity score e_i of the i-th patient via logistic regression on all of the covariates $\boldsymbol{x}_i$ with the ridge penalty for regularization, which we tune using 5-fold cross validation. The empirical densities of the propensity scores for both treatment groups are presented in Figure 2. As we can see, we have a reasonably high overlap for individuals with scores between 0.40 and 0.75. We will later on refer to this high overlap region in sensitivity analysis.

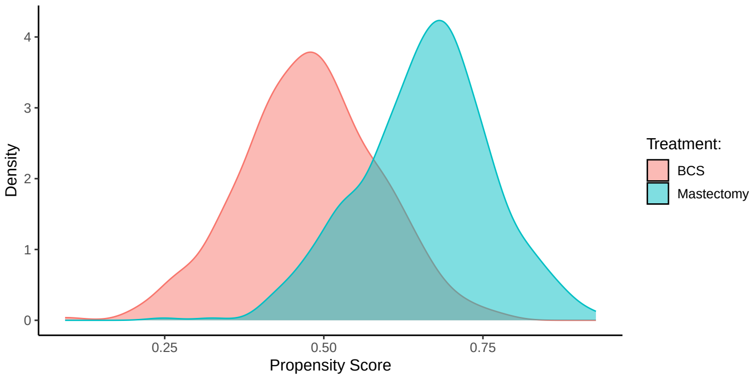

Figure 2: Propensity score density for patients depending on treatment.

Fitting the Cox model penalized with elastic net on the entire dataset allows us to perform variable selection and reduce the dimensionality of the problem. We choose the 51 predictors out of the 1363 in the original model which have nonzero values in the fit. Since we are interested in assessing the treatment effects for the entire population we ensured that the main effects term for the treatment would be among the selected predictors by not penalizing its corresponding coefficient. This way we can fit an unpenalized version of the model (2) where the vector of predictors $\boldsymbol{w}_i$ now only includes the 51 selected predictors. When fitting the model each individual i has weight given by

$$\omega_i = \frac{1}{z_i\hat{e}_i + (1-z_i)(1-\hat{e}_i)}$$

where $\hat{e}_i$ is the estimated propensity score for that individual.

4 Results

4.1 Model Fit and Diagnostics

Figure 4 in the Appendix shows the Cox-Snell residuals of the fitted model. As we can see, throughout most of the plot the $x = y$ line lies within the confidence band which indicates that the proportional hazards assumption holds reasonably well for this problem.

We present the estimated coefficients together with the corresponding 95% confidence intervals in Figure 5 in the Appendix. The estimated coefficient corresponding to the main effects of treatment is 0.261 which after exponentiation gives us 1.30, which means that we estimate that undergoing mastectomy rather than BCS increases the hazard by a factor of 1.30. However, the 95%-condifidence interval of $(-0.047, 0.569)$ includes zero so we do not have conclusive statistical evidence that choosing mastectomy over BCS increases the hazard for the patients. We find statisically significant heterogeneity in the treatment outcomes depending on the expression levels of `ush2a`, `inha`, `mapt`, and `cyp11a1` genes, with higher levels of expression of `ush2a` causing worse performance of mastectomy and higher levels of expression of `inha`, `mapt`, and `cyp11a1` resulting in improved performance of mastectomy with respect to BCS. We also find that higher levels of expression of `prkacg`, `prkg1`, the interaction between `kdm3a` and its expression level, the tumor belonging to integrative cluster 5 and higher number of lymph nodes examined positive for the presence of cancer are associated with increased hazard for patients undergoing either treatment. The large value of the estimated parameter for the `kdm3a` mutation is likely due to the fact that this mutation is rare among BIDC patients, occuring at less than 2% of the population examined here which can lead to large variance of the estimates. Meanwhile, the interaction of the presence of mutation and a high level of expression for `tbx3`, `gata3`, `tb53` as well as high expression level of `ugt2b17`, `cul1`, `e2f1`, and `brca2` are associated with a decrease in hazard across both groups of patients.

As mentioned in the Exploratory Data Analysis section, we perform sensitivity analysis with regards to the data imputation. For each of the five instances of imputed data we fit the Cox model (2) penalized via elastic net with parameter $\alpha = 0.5$ using 5-fold cross validation and record which predictors have estimated nonzero parameters corresponding to them. We then examine which subset of parameters was selected by models fitted to all the instances of imputed data. That intersection set contains 25 predictors, all of which have the same sign of the estimated coefficient. Among the 16 statistically significant parameters of the model, 9 are present in this intersection set. They are: the number of lymph nodes examined positive, tumor belonging to integrative cluster 5, `prkg1` expression level, `ugt2b17` expression level, interaction of `tp53` expression level with `tp53` mutation, interaction of `gata3` expression level with `gata3` mutation, `cul1` expression level, as well as the interaction of treatment with `cyp11a1` expression level and treatment with `mapt` expression level.

4.2 RACE Estimation

We estimate Δ_{RACE} defined in (1) by substituting S_0 and S_1 with, respectively, $\hat{S}_0$ and $\hat{S}_1$ calculated by imputing the treatment for all the individuals in the dataset and computing the estimated survival functions based on the fitted Cox model. We estimate the survival functions using the `survival` package. This gives us the estimate $\hat{\Delta}_{\text{RACE}}(t^*) = -15.7$ with the 95% bootstrap confidence interval of $(-35.1, 3.4)$. This suggests that breast conserving surgery might be more effective at ensuring longer survival, although this is not a statistically significant result.

We perform sensitivity analysis to assess the influence of patients form low-overlap regions of covariate space by considering only the subset of individuals with propensity scores from the high overlap region of 0.40 to 0.75. Based on that subset we obtain the estimate of $\hat{\Delta}_{\text{RACE}}(t^*) = -21.6$. This is reasonably close to the estimate obtained from the full dataset and lies well within the estimated confidence interval.

5 Discussion

In this study, we analyzed the effectiveness of mastectomy and breast-conserving surgery (BCS) based on the survival times of breast invasive ductal carcinoma (BIDC) patients who underwent one of these treatments, while accounting for confounders such as patient characteristics, tumor characteristics, other components of the treatment, and patients' genomes. Our goal was to estimate the restricted average survival causal effect (RACE) and identify sources of potential heterogeneity in treatment outcomes between groups of patients based on their characteristics and other treatment components involved. We compared the Cox proportional hazards model and the semiparametric accelerated failure time (AFT) model in our analysis, finding that both models had similar performance. For computational efficiency reasons, we chose to proceed with the Cox model.

Our results indicate that mastectomy might have lower effectiveness compared to BCS in ensuring longer survival of the patients. However, we do not have conclusive statistical evidence for it. We found significant heterogeneity in treatment outcomes depending on the expression levels of several genes and identified other patient characteristics which have statistically significant associations with changes in hazard for the patients, irrespective of which treatment they undergo.

In this analysis, we used the elastic net penalty, which treated all the patient characteristics and genes separately, in order to address the high dimensionality of our dataset and perform variable selection. As a future direction of research, given some expert knowledge on the mechanisms of BIDC, we could use the bi-level variable selection via sparse group lasso penalization with genes grouped according to the pathways which could be important for BIDC, analogously to the analysis by Suder and Molstad. This approach could potentially lead to models which would give us better insights into the underlying biological mechanisms of this cancer and their significance for the choice of surgical treatment.

6 References

[1] Gelman, A. (2008). *Scaling Regression Inputs by Dividing by Two Standard Deviations.* Statistics in medicine. 27. 2865-73. 10.1002/sim.3107.

[2] Mao, H., Li, L., Yang, W., Shen, Y. (2018). *On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference.* Statistics in Medicine. 37. 10.1002/sim.7839.

[3] Simon, N., Friedman, J. H., Hastie, T., Tibshirani, R. (2011). *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.* Journal of Statistical Software, 39(5), 1–13. https://doi.org/10.18637/jss.v039.i05.

[4] Suder, P. M., Molstad, A. J. (2022). *Scalable algorithms for semiparametric accelerated failure time models in high dimensions.* Statistics in Medicine. 41(6):933-949. doi: 10.1002/sim.9264

[5] van Buuren, S., Groothuis-Oudshoorn, K. (2011). *mice: Multivariate Imputation by Chained Equations in R.* Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

7 Appendix

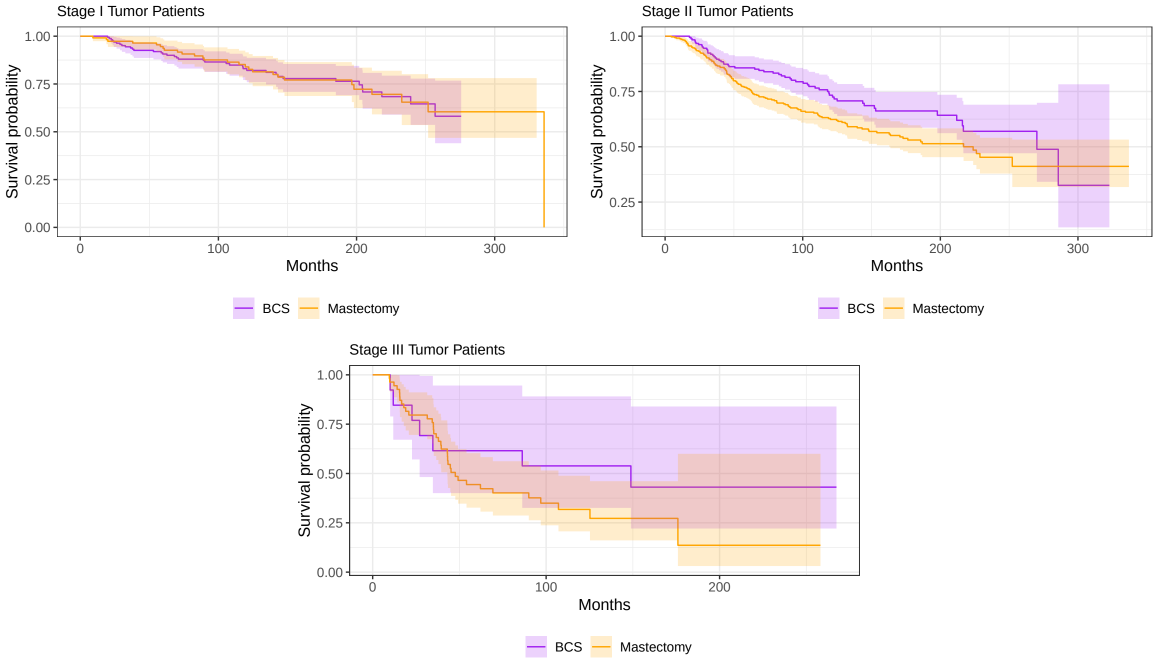

Figure 3: Survival probability over time for mastectomy and breast conserving surgery for patients with tumor stages I, II and III.

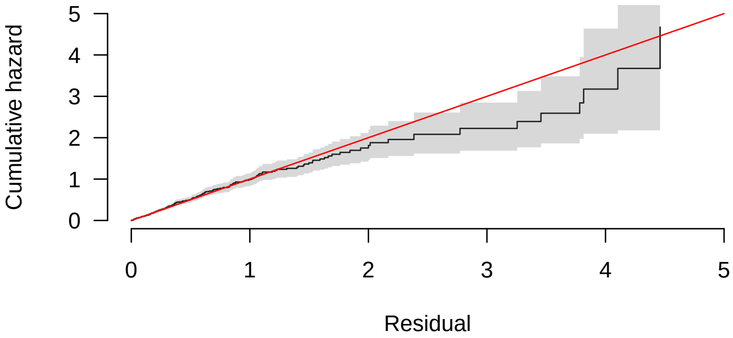

Figure 4: Cox-Snell residuals for the fitted model.

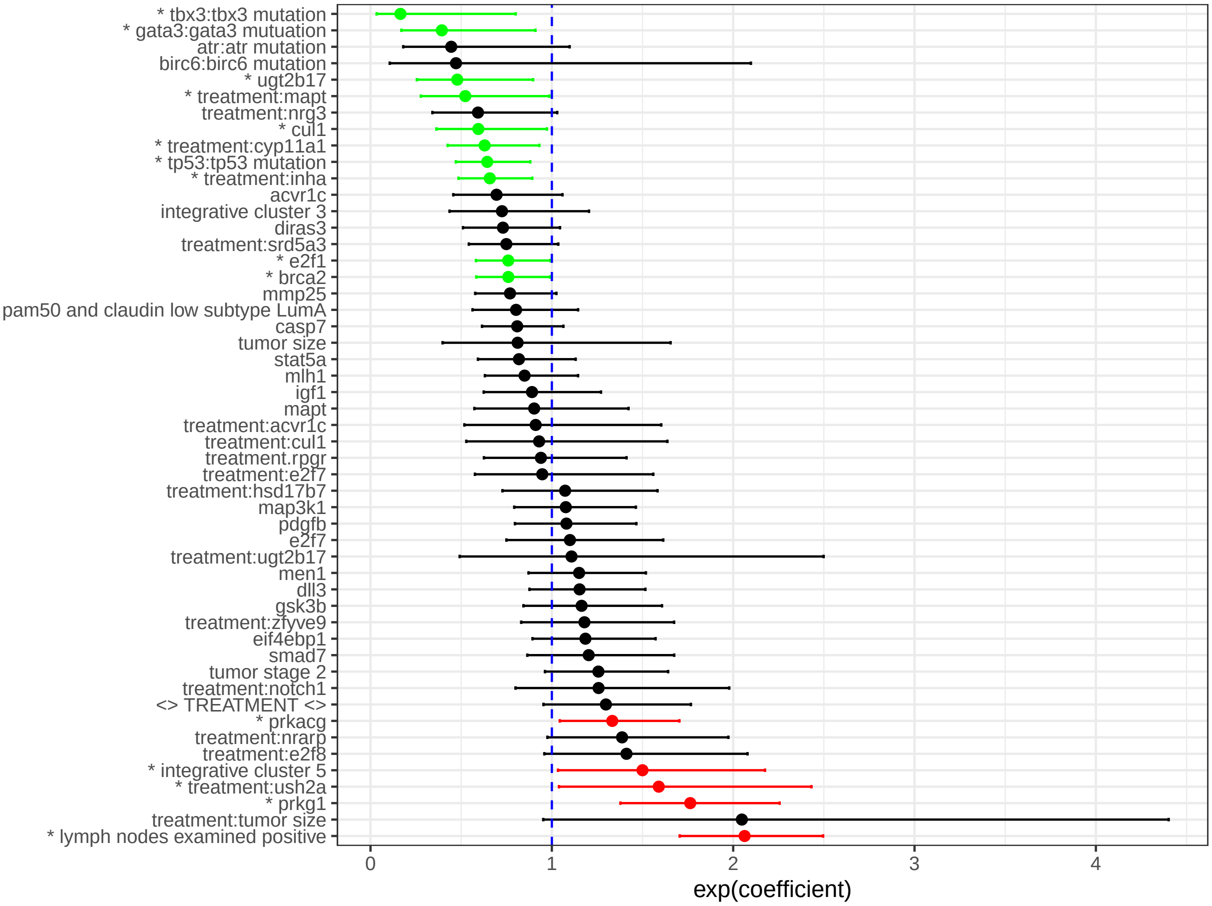

Figure 5: Estimated parameters of the Cox model. The estimates for `kdm3a:kdm3a mutation` (point estimate of 30.9 with confidence interval of (5.53, 173) after exponentiation) are not displayed due to scale difference. Such large value of the estimate is likely due to the fact that the mutation of `kdm3a` is rare among BIDC patients, occuring at less than 2% of the population examined here which can lead to large variance of the estimates.

	coefficient	estimate	exp(estimate)	robust SE	p-value
1	kdm3a.kdm3a_mut	3.43	30.9	0.878	9.38e-05
2	lymph_nodes_examined_positive	0.724	2.06	0.0972	9.21e-14
3	treatment	0.261	1.3	0.157	0.0969
4	integrative_cluster5	0.405	1.5	0.19	0.0326
5	integrative_cluster3	-0.322	0.725	0.259	0.215
6	gsk3b	0.152	1.16	0.165	0.356
7	treatment.tumor_size	0.717	2.05	0.39	0.0664
8	prkg1	0.567	1.76	0.126	6.56e-06
9	treatment.e2f8	0.345	1.41	0.198	0.0809
10	eif4ebp1	0.17	1.19	0.144	0.238
11	treatment.zfyve9	0.165	1.18	0.179	0.355
12	treatment.ush2a	0.463	1.59	0.217	0.0326
13	dll3	0.142	1.15	0.14	0.309
14	treatment.nrarp	0.327	1.39	0.18	0.0687
15	tumor_size	-0.209	0.811	0.364	0.565
16	treatment.e2f7	-0.0543	0.947	0.254	0.831
17	e2f7	0.095	1.1	0.196	0.627
18	smad7	0.185	1.2	0.169	0.272
19	men1	0.14	1.15	0.142	0.322
20	prkacg	0.288	1.33	0.125	0.0213
21	treatment.notch1	0.229	1.26	0.231	0.321
22	pdgfb	0.0773	1.08	0.156	0.619
23	tumor_stage2	0.228	1.26	0.136	0.0936
24	mmp25	-0.262	0.769	0.146	0.0732
25	treatment.acvr1c	-0.0933	0.911	0.288	0.746
26	treatment.srd5a3	-0.289	0.749	0.165	0.0802
27	treatment.hsd17b7	0.0706	1.07	0.198	0.722
28	mlh1	-0.163	0.85	0.152	0.284
29	casp7	-0.212	0.809	0.14	0.129
30	brca2	-0.275	0.76	0.135	0.042
31	treatment.rpgr	-0.0622	0.94	0.208	0.765
32	treatment.ugt2b17	0.103	1.11	0.415	0.804
33	igf1	-0.116	0.891	0.181	0.523
34	treatment.nrg3	-0.523	0.593	0.282	0.0634
35	map3k1	0.0741	1.08	0.156	0.636
36	mapt	-0.102	0.903	0.232	0.659
37	e2f1	-0.275	0.76	0.136	0.043
38	ugt2b17	-0.737	0.478	0.321	0.0215
39	treatment.inha	-0.419	0.658	0.155	0.00701
40	diras3	-0.314	0.73	0.183	0.0856
41	tp53.tp53_mut	-0.441	0.644	0.16	0.00594
42	atr.atr_mut	-0.809	0.445	0.461	0.079
43	gata3.gata3_mut	-0.933	0.393	0.427	0.029
44	acvr1c	-0.364	0.695	0.215	0.09
45	cul1	-0.52	0.595	0.251	0.0385
46	stat5a	-0.2	0.819	0.165	0.225
47	treatment.cul1	-0.0728	0.93	0.289	0.801
48	pam50_._claudin.low_subtypeLumA	-0.22	0.803	0.181	0.225
49	birc6.birc6_mut	-0.752	0.471	0.762	0.323
50	treatment.cyp11a1	-0.463	0.629	0.2	0.0204
51	tbx3.tbx3_mut	-1.8	0.165	0.805	0.0253
52	treatment.mapt	-0.649	0.523	0.324	0.0451

Table 1: Fitted Cox model parameters.