Journal of Biomedical Analytics
Vol. 1, No. 1 (2018), pp. 1–14
doi:10.30577/jba.2018.v1n1.5

Research Article

Association between Gene Expression, Clinical Factors and Survival in Patients with Breast Cancer

Tasnia Ahmed and Munni Begum

Ball State University, USA

Received: September 30, 2017; revised: January 23, 2018; accepted: January 27, 2018.

Abstract: Breast cancer is the most frequently diagnosed and the second cause of cancer deaths among women. Several genes are found to be significantly responsible for developing breast cancer. When healthy, these genes act as tumor suppressors by producing a protein that prevents cells from growing uncontrollably. But with mutations and other disorders in these genes, cells can grow quickly and tumors may form. In previous research it has been shown that higher risk of developing breast cancer is associated with having a number of gene mutation, but it is not clear how genetic factors affect the survival pattern in breast cancer patients in the presence of clinical and other relevant factors. In this paper, we consider the joint effect of a number of important genes and relevant clinical factors to study the survival pattern of breast cancer patients using data from The Cancer Genome Atlas (TCGA) project. Among the clinical variables, increasing age, premenopausal status, prior history of cancer and neoplasm status with tumor, Black or African American race and stage IIIA have significantly higher risk of failure (death) from breast cancer. We found significant difference in survival time between altered and non-altered cases in four genes FOXA1, MLH1, RAD50, and RAD51C while considered genetic factors only. After controlling for clinical factors, patients with three mutated genes RAD50, PTEN, and MAP3K1 had higher relative risk of failure from breast cancer.

Keywords: Breast Cancer, Gene expression, Clinical factors, RNA-Seq, Survival analysis.

1 Introduction

Breast cancer which starts in the tissues of the breast is the most frequently diagnosed cancer and the second cause of cancer deaths in women. Invasive Ductal Carcinoma (IDC) is the most common type of breast cancer and it comprises about 70% to 80% of all breast cancer (National Breast Cancer, 1991). About 10% of all cases of advanced breast cancer are Invasive Lobular Carcinoma (ILC). In 2017, an estimated 255,180 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S, along with 63,410 new cases of non-invasive breast cancer (Breastcancer.org, 1999). Men can also have breast cancer, although breast cancer in male is not as prevalent as in women. About 2,470 new cases of invasive breast cancer are expected to be diagnosed in men in 2017 (Breastcancer.org, 1999).

Prognostic factors of breast cancer constitute both clinical and genetic factors as found in the literature. A number of studies (Faradmal *et al.*, 2012; Abadi *et al.*, 2014; Bilal *et al.*, 2013), investigated the role of clinically associated factors such as, age at menopause, age at diagnosis, stage of the disease, tumor size, histological grade, type of therapy received (hormone therapy, radiotherapy, or chemotherapy), and family history on patient's survival. Similarly a number of studies (Martin and Weber, 2000; Apostolou and Fostira, 2013; Ciriello *et al.*, 2015), addressed association of genes and their mutations with occurrences of breast cancer. In order to investigate the joint role of clinical and genetic factors on the breast cancer patient's survival we abstracted survival, clinical, and gene expression data from TCGA separately and combined them.

Gene expression levels can provide how information from a gene is used in the composition of a functional gene product (Wikipedia, 2001a). Products are often proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA. Expressions are quantified to study cellular changes in response to external changes or stimuli and to study differences between healthy and diseased states.

RNA-Seq (RNA Sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given time (Wikipedia, 2001b). It is generally used to compare differential gene expressions between two or conditions, such as treated vs non-treated, wild-type versus mutant, and to find out which genes are up- or down-regulated in each condition. We considered the RNA-Seq expression data for all the patients as genetic factors from TCGA database.

In previous research it has been shown that higher risk of developing both invasive breast cancer is associated with having a number of gene mutation. One study based TCGA data, (TCGA, 2012) showed mutations in number of genes are associated with many subtypes of breast cancer. Another study (Giovanni *et al.*, 2015), profiled both IDC and ILC using TCGA data and found association between mutations in a number of genes and development of IDC and ILC. Although there was attempt to include both clinical factors and genetic profiling (Haoming *et al.*, 2015), in identifying the risk of developing breast cancer, it is not clear how genetic factors affect the relative risk of developing breast cancer and patients' survival in the presence of clinical and other relevant factors. In this paper, we consider the joint effect of a number of important genes and other relevant factors to study the survival pattern of breast cancer patients using data from The Cancer Genome Atlas (TCGA) project.

2 Data and Variables

2.1 Data Source

The RNA-Seq data for this study has been retrieved from the Open Access data tier of TCGA genome data analysis center (http://gdac.broadinstitute.org/) which is an interactive data system for researchers to search, download, upload, and analyze cancer genomic data, including breast cancer data (TCGA, 2012). Since we wanted to explore a particular point of interest, survival analysis of breast cancer on clinical and genetic factors, we retrieved the clinical data (Breast Invasive Carcinoma TCGA, Provisional) from cBioPortal which is an exploratory analysis tool for exploring large-scale cancer genomic data sets.

2.2 The Breast Cancer Dataset

The clinical data contains 1105 cases and 118 variables whereas the RNA-Seq data with gene expression information has 1100 cases with 17675 gene information. We considered 9 clinical factors and 25 genes (Antoniou and Easton, 2006; Easton, 1999), among all the available information, which were found to be stated as significant factors (Cerami *et al.*, 2012; Yang *et al.*, 2007) in the breast cancer literature. A single outcome, i.e breast cancer specific survival is investigated for the survival analysis. Since we need the same number of patients in both clinical and RNA-Seq data, we matched the patient ID in both data and found 1100 patients that fulfill the matching criterion. Two hundred and ninety one patient records with unknown clinical information have been excluded from downstream analysis for the remaining 809 cases.

Among patient information gender, menopause status (at time of diagnosis), race, tumor status (at time of last contact or death), vital status (at date of last contact) are chosen (Helmrich *et al.*, 1983). Histologic subtype and age at initial diagnosis are selected from pathologic information and stage variable comes from (AJCC) staging manual (Giuliano *et al.*, 2017). In order to create the survival object (response variable), information were collected on number of days from the date of initial pathologic diagnosis to date of death and date of last contact for patients who were initially diagnosed as breast cancer carrier between year 1988 and 2013.

We identified normal and tumor samples by using the TCGA barcode. The two digits at position 14-15 of the barcode indicates the sample type. Tumor type ranges from 01-09, normal type from 10-19, and control samples from 20-29. In gene expression analysis, a common approach to differentiate a gene's expression between two conditions is to use a threshold value for what is called *fold change(FC)* (Tutorial, 2015). Simply considering *FC* as a measure of differential expression may not be valid in this case. Thus we scale the gene expression value by a standardized transformation and calculate z-score for each expression value.

For gene expression data, the standard rule is to compute the relative expression of an individual gene and tumor to the gene's expression distribution in a reference population. The reference population is either all tumors that are diploid (containing two complete sets of chromosomes, one from each parent) for the gene in question or when available, normal adjacent tissue. The returned value indicates the number of standard deviations away from the mean of expression in the reference population (z-score). This measure is useful to determine whether a gene is up- or down-regulated relative to the normal samples or all

other tumor samples. In this study, we considered those genes with $z > \pm 1.96$ (roughly $p = 0.05$ or 2 SD away) to be differentially expressed. To obtain z-scores for the RNA-Seq data we use following formula:

$$z = \frac{\text{Expression for gene } X \text{ in tumor } Y - \text{Mean expression for gene } X \text{ in normal}}{\text{Standard deviation of expression for gene } X \text{ in normal}}$$

We used the z-score values to define which samples are altered and which do not change. The numbers for *altered* and *not altered* refer to the number of samples with gene expression higher/lower than a specific threshold such as z-score of 2. That is we can define *altered* versus *not altered* as follows:

$$z \geq 2 \Rightarrow altered$$

$$z < 2 \Rightarrow \textit{not altered}$$

The dependent variable for this study is overall survival time after the diseases has been initially diagnosed. To perform survival analysis we need the following three main constructs:

- time: the time till an event happens
- status: indicates which patients have to be kept for the analysis
- event: indicates which patients have death after initial pathological diagnosis (IPD).

In addition to these three constructs, we also define a censoring indicator as follows:

$$\text{Censoring indicator} = \begin{cases} 1; \text{ if a patient is alive at time of survey} \\ 0; \text{ otherwise.} \end{cases}$$

The time variable is defined as the number of days to death after IPD and number of days to last contact after IPD (for censored cased).

For our analysis data, the median survival time of the breast cancer patients is 3472 days after IPD with 87.38% censoring in the data.

3 Methods

To determine the moderated effect of important clinical factors of breast cancer on patient's survival with information on gene expression for a number of significant genes, we fit the standard Cox Proportional Hazards (Cox, 1972) model for univariate analysis and the penalized Cox PH model (Heinze and Schemper, 2001) for combined analysis. Important clinical and demographic factors associated to the disease are selected consulting the literature on breast cancer. We also search the literature to narrow down a short list of genes found so far to be linked to Invasive Ductal Carcinoma (IDC). Twenty five genes that are found to be directly associated with IDC are selected. For each gene its expression value is converted to z-score and based on a threshold value, as discussed in section 2, each gene is labeled as *altered* versus *not altered*. To further narrow down the number of genes that will be added to the combined Cox PH model, we considered both univariate and multivariate analysis for each gene alone, and with rest of the genes excluding the gene under investigation respectively. Survival functions for the *altered* and *not altered* groups are compared under both scenarios. If the survival functions are statistically significantly different among the altered

versus not altered groups for a particular gene, then it is added to the combined model for further analysis. An ultra-brief overview of survival analysis is presented as follows.

A subject or patient's survival status can be estimated by calculating what is called the Product-limit (PL) estimator (Kaplan and Meier, 1958) of the survival function defined as $S(\hat{t}_j) = \prod_{i=1}^{j} \left(1 - \frac{d_j}{n_j}\right)$. Here $S(\hat{t}_j)$ is the estimated survival function at time t_j, d_j is the number of events occurred at t_j, and n_j is the number of subjects available at t_j. After estimating survival functions, we can compare these among two or more groups using *Log-rank test* (Mantel, 1966). For instance, we implemented *Log-rank test* to identify the most important genes whose status of being altered or not are associated significantly with patient survival. The null hypothesis for this test can be formulated as follows:

H_0 : Survival function for patients with altered gene is the same as that for patients with non-altered gene

H_A : Survival functions are not equal for these two groups

Symbolically we can write,

$$H_0 : S_{\text{altered}}(t) = S_{\text{not altered}}(t)$$

$$H_A : S_{\text{altered}}(t) \neq S_{\text{not altered}}(t)$$

If the survival function is significantly different among altered and not altered groups for a specific gene, we select and include it to the combined Cox PH model. This approach is useful in learning the effect of this specific gene under investigation on patient's survival in the presence of clinical factor and other selected genes.

The semi-parametric Cox PH model (Cox, 1972) is commonly used regression model for time-to-event response variable. Cox PH model fits the hazard of having the event under consideration (IDC, in this case), using an unspecified baseline hazard function and an exponentiated form of a set of covariates. Mathematically the model can be written as follows:

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

where $h(t|\mathbf{x})$ is the conditional hazard function for a subject i with covariate information given as the vector $\mathbf{x}_i$, $h_0(t)$ is the baseline hazard function that is independent of covariate information, and $\boldsymbol{\beta}$ is the vector of regression coefficients corresponding to the covariates.

In order to determine if a particular predictor has any effect on patient's survival, we calculate what is called the *hazard ratio (HR)* based on the estimated regression coefficients from the fitted Cox PH model. The hazard ratio for a covariate x_r can be expressed by the following simple formula e^{β_r}. Thus hazard ratio for any covariate can be obtained by exponentiating the corresponding regression coefficient.

4 Results and Discussions

As mentioned in section 2, we considered nine clinical factors including age and twenty five genes in our analysis. Except age, rest of the clinical factors are categorical. Patient's age distribution is shown in Figure 1 below. We see that the mean age of the patients at the time of diagnosis is approximately 58 (58.34) years with a minimum of 26 and maximum of 90

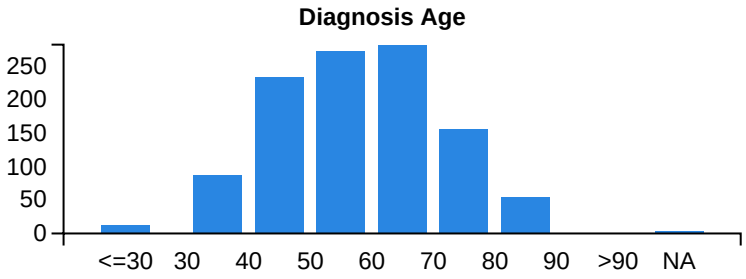

Figure 1: Age of the patients at the time of diagnosis.

years. Note that, there are 2 missing observations in age variable which is indicated by NA in the graph.

We present the descriptive summary statistics of the other clinical predictors in Table 1.

Table 1: Descriptive statistics of clinical predictors.

Characteristics	**Category**	**Freq**	**%**
Race	White	759	69.00
	Black/African American	183	16.64
	Other[1]	62	5.64
Prior History	Yes	67	6.09
	No	1031	93.73
Neoplasm Status	With Tumor	95	8.64
	Tumor Free	879	79.91
Stage	Stage I	90	8.18
	Stage IA	85	7.73
	Stage IB	6	0.55
	Stage II	6	0.55
	Stage IIA	358	32.55
	Stage IIB	260	23.64
	Stage III	2	0.18
	Stage IIIA	155	14.09
	Stage IIIB	27	2.45
	Stage IIIC	67	6.09
	Stage IV	20	1.82
	Stage X	12	1.09
Menopause Status[2]	Premenopausal	230	20.91
	Perimenopausal	40	3.64
	Postmenopausal	704	64.00
	Indeterminate	34	3.09

[1]Other: Asian, American Indian or Alaska native.

[2]Premenopausal:<6 months since last menstrual period (LMP) and no prior bilateral oophorectomy and not on estrogen replacement.
Perimenopausal: 6-12 months since last menstrual period.

From Table 1, we see that the stage variable has the highest number (12) of categories and it is found that 32.55% patients belong to stage IIA, while 23.64% and 14.09% patients are from stage IIB and IIIA respectively. The percentages for other categories are relatively small than these three categories.

Proportion of white patients is the highest with 69%, 16.64% of the patients are Black or African American, and only 5.64% patients are Asian and American Indian or Alaska native. There are four categories for menopause status. In addition, 64% patients have post-menopausal, 20.91% of the patients have premenopausal status, 3.64% is perimenopausal, and 3.09% of the patients is indeterminate. In the original dataset, there were eight categories for histology type. Among these categories, 71.55% patients have Infiltrating Ductal Carcinoma, 18.45% have Infiltrating Lobular Carcinoma and only 10% of the patients have other types of histology.

4.1 Survival Pattern for Gene Expression Data

Using product limit (PL) estimator, we obtained survival curves for each of the twenty-five genes. The survival curves compare survival patterns between two groups: *altered* and *non-altered*. Here we included the results from the genes those have significant difference in their survival pattern. The p-value indicates the significant role of the gene in differential survival pattern when its expression level is considered as *altered* versus *not altered*.

Note that, in Figure 2, the red line in the graphs indicates non altered and the black line indicates for altered group. We see that patients who have altered MLH1, FOXA1 and RAD51C genes are less likely to survive compared to the non-altered groups among these genes. However, p-value for RAD50 gene indicates that patients with alteration of this gene have a better survival than non-altered group.

4.2 Fitting Penalized Cox PH Model

To select the most significant genes for the final analysis we used penalized Cox PH model (Heinze and Schemper, 2001) by penalizing all twenty-five genes keeping the clinical factors unpenalized. We considered a range of values for the shrinkage or penalty parameter λ from 3 to 20. Based on the non zero regression coefficients, we selected 13 genes from the RNA-Seq data for the final model. Table 2 and Table 3 show the estimated coefficients (β), corresponding hazard ratios (HR), z, and p-values from these analyses considering $\lambda = 20$ and $\lambda = 15$ respectively.

While the shrinkage parameter value λ is held at a higher value of 20, we see from Table 2 that the coefficients for genes BRCA1, GATA3, PIK3CA, RAD50, and RAD51C were not shrunk toward zero. This implies that these five genes are important while controlling for the clinical factors, although only RAD50 is statistically significant at the 10% level of significance.

Table 3 presents the results when we reduce the shrinkage parameter value λ from 20 to 15. We see from that the coefficients for four additional genes BARD1, MLH1, MSH2, and PTEN were not shrunk toward zero. This implies that these nine genes are important while controlling for the clinical factors, again only RAD50 is statistically significant at the 10% level of significance. In addition it is to be noted that BRCA1, GATA3, PIK3CA, RAD50,

Postmenopausal: Prior bilateral ovariectomy or $>$ 12 months since LMP with no prior hysterectomy.
Indeterminate: Unknown.

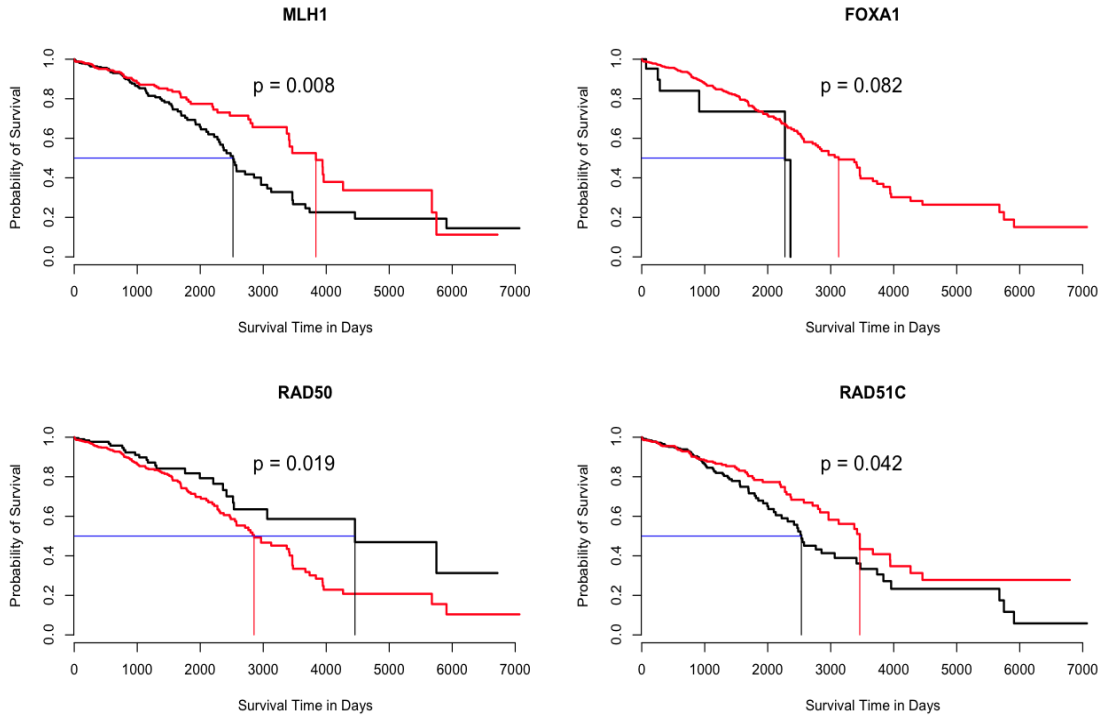

Figure 2: Survival pattern of *altered* and *non-altered* groups of MLH1, FOXA1, RAD50 and RAD51C.

Table 2: Clinical and genetic factors from combined penalized analysis with penalty parameter =20.

Variables	β	HR	z	p
Age	0.6077	1.05	4.02	**0.0391**
White	-1.67	2.34	-1.78	**0.075**
Other	1.83	1.64	0.65	0.519
Peri-menopausal	-1.07	2.75	-0.27	0.490
Post-menopausal	-1.03	2.89	-0.69	0.2356
Indeterminate	-1.23	1.07	-2.27	**0.023**
Prior History (Yes)	3.37	2.50	2.16	**0.030**
Neoplasm Status (With Tumor)	1.95	7.01	7.88	$\mathbf{9.79}e^{-07}$
StageI	-3.05	1.60	-1.82	**0.069**
Stage IA	-1.48	0.84	-1.94	**0.053**
Stage IB	-4.43	0.0027	0.00	0.997
Stage II	-4.16	0.0020	-0.00	0.997
Stage IIA	-0.95	2.84	-0.77	0.444
Stage IIB	-3.27	1.51	-2.50	**0.012**
Stage IIIB	-0.65	3.09	-0.33	0.744
Stage IIIC	0.86	1.26	0.50	0.620
Stage IV	-0.51	3.54	-0.08	0.938
Stage X	1.02	2.76	1.45	0.146
BRCA1 (Not Altered)	-0.86	3.45	-0.27	0.786
GATA3 (Not Altered)	-0.32	3.59	-0.07	0.941
PIK3CA (Not Altered)	1.21	1.09	0.38	0.700
RAD50 (Not Altered)	2.01	1.73	1.80	**0.072**
RAD51C (Not Altered)	-0.91	2.87	-1.06	0.291

and RAD51C are retained by both penalized models. In our final combined analysis, we considered the genes that are retained with $\lambda = 10$ in order to keep higher number of influential genes. Finally we identified important genes for breast cancer controlling for the clinical factors using the regular Cox model.

4.3 Modeling Hazard on the Combined Model with both Clinical and RNA-Seq Data

We included seven out of nine clinical factors identified from the separate analysis on clinical factors and thirteen out of twenty five genes identified using penalized Cox model to our combined analysis.

Estimated regression coefficients are presented in Table 4 along with hazard ratios (HR), z-values, and corresponding p-values. Here, n= 808, number of events(deaths)= 102 after deleting 292 observations due to missingness.

Table 3: Clinical and genetic factors from combined penalized analysis with penalty parameter =15.

Variables	β	HR	z	p
Age	0.67	1.05	4.30	**0.0115**
White	-1.66	2.34	-1.72	**0.085**
Other	1.93	1.69	0.68	0.497
Peri-menopausal	-0.90	2.88	-0.23	0.490
Post-menopausal	-1.55	2.41	-1.02	0.306
Indeterminate	-1.39	0.92	-2.47	**0.014**
Prior History (Yes)	3.54	2.62	2.26	**0.024**
Neoplasm Status (With Tumor)	1.99	7.35	7.62	$\mathbf{2.4}e^{-14}$
StageI	-3.06	1.60	-1.84	**0.066**
Stage IA	-1.53	0.80	-1.99	**0.047**
Stage IB	-4.24	0.0016	0.00	0.996
Stage II	-3.89	0.0055	-0.01	0.995
Stage IIA	-1.42	2.50	-1.13	0.257
Stage IIB	-1.06	1.28	-2.88	**0.004**
Stage IIIB	-0.56	3.16	-0.28	0.777
Stage IIIC	0.49	1.14	0.28	0.782
Stage IV	-1.42	2.50	-0.75	0.453
Stage X	3.20	2.39	1.20	0.230
BRCA1 (Not Altered)	-0.60	3.12	-0.67	0.505
BARD1 (Not Altered)	1.58	1.54	1.23	0.220
GATA3 (Not Altered)	-0.15	3.64	-0.04	0.972
MLH1 (Not Altered)	-0.73	3.02	-0.87	0.386
MSH2 (Not Altered)	1.15	1.09	0.27	0.786
PTEN (Not Altered)	-1.32	2.57	-1.39	**0.164**
PIK3CA (Not Altered)	1.30	1.10	0.41	0.682
RAD50 (Not Altered)	2.31	1.87	2.02	**0.044**
RAD51C (Not Altered)	-0.96	2.84	-1.11	0.265

Note that, Black or African American group has been used as the reference category for race variable. Similarly, the references for menopause status, stage, prior history, and neoplasm status are pre-menopausal group, stage IIIA, patients with prior occurrence of breast cancer, and with tumor respectively. Also, the altered group has been considered as the reference for all the selected genes. From Table 4, we see that there is 1.05 times increase in the expected hazard relative to a one year increase in age. The patients who belong to White race category have 2.36 times less failure rate compared to the African Americans or Black patients and have significant effect on death from breast cancer. Others have 1.56 times more failure rate than African Americans but does not have significant effect on overall survival status.

In the case of prior history, we can see that patients who have history of prior disease occurrences are 2.74 times more likely to die compared to a patient with no prior history. This factor shows significant effect on death from breast cancer.

Table 4: Clinical and genetic factors from the combined analysis.

Variables	β	HR	z	p
Age	0.71	1.05	4.50	**0.0166**
White	-1.63	2.36	-1.64	**0.1008**
Other	1.64	1.56	0.57	0.5658
Peri-menopausal	-0.85	2.92	-0.22	0.8262
Post-menopausal	-1.82	2.24	-1.19	0.2356
Indeterminate	-1.30	1.00	-2.30	**0.0213**
Prior History (Yes)	1.01	2.74	2.36	**0.0185**
Neoplasm Status (With Tumor)	2.08	7.97	7.94	$\mathbf{6.12}e^{-07}$
StageI	-3.33	1.49	-1.99	**0.0463**
Stage IA	-1.69	0.68	-2.19	**0.0284**
Stage IB	-4.27	$1.46e^{-07}$	0.00	0.9964
Stage II	-3.89	$6.04e^{-07}$	-0.01	0.9953
Stage IIA	-1.82	2.24	-1.41	0.1574
Stage IIB	-1.21	1.10	-3.22	**0.0013**
Stage IIIB	-0.16	3.64	-0.02	0.9831
Stage IIIC	0.42	1.12	0.24	0.8075
Stage IV	-2.41	1.91	-1.25	0.2128
Stage X	2.35	1.89	0.88	0.3785
BRCA1 (Not Altered)	-0.71	3.04	-0.76	0.4479
BARD1 (Not Altered)	0.58	1.17	0.44	0.6593
GATA3 (Not Altered)	-0.83	3.46	-0.19	0.8506
MLH1 (Not Altered)	-0.95	2.84	-1.10	0.2703
MSH2 (Not Altered)	0.41	1.03	0.10	0.9229
MRE11A (Not Altered)	0.59	1.17	0.45	0.6512
MAP3K1 (Not Altered)	1.11	3.02	1.86	**0.0634**
NBN (Not Altered)	0.68	1.20	0.69	0.4922
PTEN (Not Altered)	-1.73	2.30	-1.76	**0.0783**
PIK3CA (Not Altered)	0.66	1.05	0.18	0.8597
RUNX1 (Not Altered)	-0.43	3.56	-0.04	0.9670
RAD50 (Not Altered)	2.59	2.02	2.21	**0.0268**
RAD51C (Not Altered)	-0.94	2.85	-1.08	0.2803

Tumor status is a vital factor for death from any cancer. From Table 4, we can see that patients with tumor have 7.97 times more failure rate compared to a patient with no neoplasm tumor and is highly significant on overall survival.

Stage level plays an important role on death from breast cancer. From the Table 4, it can be seen that patients who belong to Stage I and IIB have respectively 1.49 and 1.10 times less failure rate compared to a patient from Stage IIIA and these two stage levels have significant effect on death. Moreover, patients who belong to stage IA are 0.68 times less likely to die compared to the patient at stage IIIA and is statistically significant.

After controlling for the clinical factors, only MAP3K1, PTEN, and RAD50 show significant influence on the survival of breast cancer. The patients with non-altered MAP3K1 and RAD50 genes have respectively 3.02 and 2.02 times more rates of death compared to the

altered carriers. Whereas, patients with non-altered PTEN have 2.30 times less rate of death than the altered carriers.

5 Conclusion

In this study, a survival analysis of 1100 breast cancer cases revealed that a number of gene mutations independently predicts breast cancer survival, whereas some of them were not significantly associated with overall higher survival.

From product limit (also known as Kaplan Meier) analysis, we found significant difference in survival times between altered and non-altered cases in four genes namely, FOXA1, MLH1, RAD50, and RAD51C. After controlling for clinical factors, three RAD50, PTEN, and MAP3K1 mutated patients had higher relative risk of failure from breast cancer.

Furthermore, using combined data for both clinical variables and gene expression from the same 1100 breast cancer cases, we identified the most significant factors to predict overall survival. Similar to the results of univariate analysis from Kaplan-Meier estimates, RAD50 showed significant association with survival from breast cancer. In addition, MAP3K1 and PTEN genes had been identified as significant predictors for survival where the altered carriers of PTEN had worse survival or higher risk of death than the non-altered cases after controlling for the clinical variables. The patients with altered MAP3K1 had better survival than the non-altered carriers. Our results on the influentials genes for development of breast cancer after controlling for clinical factors conform with findings from previous studies (TCGA, 2012; Giovanni *et al.*, 2015).

While we conducted the analysis on clinical variables, controlling for the genetic factors, patients with increasing age, pre-menopausal status, prior history of cancer and neoplasm status with tumor have significantly higher risk of failure (death) from breast cancer. Also, patients who belong to White race group have lower rate of death compared to Black or African American. Patients at stage I, IA and IIB have significantly better survival than patients from at IIIA.

Although, to our knowledge, the patient cohort (1105 cases) includes nearly all important factors (both genomic and clinical) combined, it is possible that some clinical and genetics factors are not recorded and our findings should be further validated with integrated platforms. Another shortcoming of this research is that we limited to twenty five genes based on the literature search on breast cancer. An objective gene selection from all available genes in the database is left as future research.

Missingness is a major drawback in biomedical studies. In particular, often times missing data fail to observe and include all important information about the patients. We completely ignored the missingness of the covariates and excluded from all downstream analysis. Taking into account the covariates for which information was incomplete might give us more accurate results.

The results should be interpreted with caution as we have not cross-validated the findings across other platforms, such as genomic DNA copy number arrays, DNA methylation, or protein arrays (TCGA, 2012). Nevertheless, the discovery of combined effect of important clinical and genomic predictors on patient survival in breast cancer may have important implications in identifying influential genes controlling for clinical factors and vice versa.

Declarations

Acknowledgments

The authors would like to thank three anonymous referees for their critical readings and most helpful comments which improve the paper significantly.

Funding

None.

Conflict of Interest

The authors declare that they have no competing interests.

Ethical Approval

Not applicable.

References

Abadi A, Yavari P, Dehghani-Arani M, Alavi-Majd H, Ghasemi E, Amanpour F, Bajdik C (2014). "Cox models survival analysis based on breast cancer treatments." *Iranian Journal of Cancer Prevention*, **7**(3), 124–129.

Antoniou AC, Easton DF (2006). "Models of genetic susceptibility to breast cancer." *Oncogene*, **25**(43), 5898–5905. doi:10.1038/sj.onc.1209879.

Apostolou P, Fostira F (2013). "Hereditary breast cancer: the era of new susceptibility genes." *BioMed Research International*. doi:10.1155/2013/747318.

Bilal E, Dutkowski Jand Guinney J, Jang IS, Logsdon BA, Pandey Gea (2013). "Improving breast cancer survival analysis through competition-based multidimensional modeling." *PLoS Comput Biol*, **9**(5). doi:10.1371/journal.pcbi.1003047.

Breastcancerorg (1999). "Breast Cancer Fact Sheet." http://www.breastcancer.org/about_us/press_room/press_kit/facts_figures. Last Accessed Nov 29, 2017.

Cerami Eand Gao J, Dogrusoz Uand Gross BE, Sumer SO, Aksoy BAea (2012). "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data." *Cancer Discov.*, **2**(5), 401–404. doi:10.1158/2159-8290.CD-12-0095.

Ciriello G, Gatza ML, Beck AH, Wilkerson M Dand Rhie SK, Pastore Aea (2015). "Comprehensive molecular portraits of invasive lobular breast cancer." *Cell*, **163**(2), 506–519. doi:10.1016/j.cell.2015.09.033.

Cox D (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2), 187–220.

Easton DF (1999). "How many more breast cancer predisposition genes are there?" *Breast Cancer Research*, **1**(1). doi:10.1186/bcr6.

Faradmal J, Talebi A, Rezaianzadeh A, Mahjub (2012). "Survival analysis of breast cancer patients using cox and frailty models." *Journal of Research in Health Sciences*, **12**(2), 127–130.

Giovanni C, Michael LG, Andrew HB, Matthew DW, Suhn KRea (2015). "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer." *Cell*, **163**(2), 506âĂŞ–519. doi:10.1016/j.cell.2015.09.033.

Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, Weaver DL, Winchester DJ, Hortobagyi GN (2017). "Breast Cancer-Major changes in the American Joint Committee on Cancer." *CA: A Cancer Journal for Clinicians*, **67**, 290–303. doi:10.3322/caac.21393.

Haoming X, Mohammad AM, Pietro L (2015). "Network regularised Cox regression and multiplex network models to predict disease comorbidities and survival of cancer." *Computational Biology and Chemistry*, **59**, 15–31. doi:10.1016/j.compbiolchem.2015.08.010.

Heinze G, Schemper M (2001). "A solution to the problem of monotone likelihood in Cox regression." *Biometrics*, **57**(1), 114–119. doi:10.1111/j.0006-341x.2001.00114.x.

Helmrich SP, Shapiro S, Rosenberg L, Kaufman DW, Slone D, Bain Cea (1983). "Risk factors for breast cancer." *American journal of epidemiology*, **117**(1), 35–45. doi:10.1093/oxfordjournals.aje.a113513.

Kaplan E, Meier P (1958). "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association*, **53**(282), 457–481. doi:10.2307/2281868.

Mantel N (1966). "Evaluation of survival data and two new rank order statistics arising in its consideration." *Cancer Chemother. Rep.*, **50**, 163–170.

Martin AM, Weber BL (2000). "Genetic and hormonal risk factors in breast cancer." *Journal of the National Cancer Institute*, **92**(14), 1126–1135.

National Breast Cancer (1991). "The National Breast Cancer Foundation." http://www.nationalbreastcancer.org/invasive-ductal-carcinoma. Last Accessed Dec 31, 2017.

TCGA (2012). "Comprehensive molecular portraits of human breast tumours." *Nature*, **490**, 61–70. doi:10.1038/nature11412.

Tutorial (2015). "Survival analysis of TCGA patients integrating gene expression (RNASeq) data." https://www.biostars.org/p/153013/#179081. Last Accessed Sep 30, 2017.

Wikipedia (2001a). "Gene expression." https://en.wikipedia.org/w/index.php?title=Gene_expression&oldid=802897631. Last Accessed Sep 30, 2017.

Wikipedia (2001b). "RNA-Seq." https://en.wikipedia.org/w/index.php?title=RNA-Seq&oldid=766559087. Last Accessed Apr 7, 2017.

Yang XR, Sherman ME, Rimm DL, Lissowska J, Brinton LA, Peplonska B, Hewitt SM, Anderson WF, Szeszenia-Dąbrowska N, Bardin-Mikolajczak A, *et al.* (2007). "Differences in risk factors for breast cancer molecular subtypes in a population-based study." *Cancer Epidemiology Biomarkers Prev.*, **16**(3), 439–443. doi:10.1158/1055-9965.EPI-06-0806.