A structural equation modeling approach for the identification of gene expression adaptation in breast cancer patients at grade specific level

Rahul V. Veettil and Norm O'Rourke

Ben -Gurion University of the Negev

Abstract

Introduction: Breast cancer is a common cancer in women, but screening and therapy has improved the survival rates dramatically. However, recurrence remains to be a major problem faced by patients. Most drugs become ineffective when a cancer reappear, due to the aggressive growth rate of recurrent tumor. Effective classification of tumor grade before therapy could help the patients to avoid recurrence and allow clinicians to choose optimal therapy.

Methods: We used classical machine learning approach information gain and structural equation modeling to prioratize and find structural relationship between breast cancer genes, recurrence events and recurrence free survival using AMOS software in a large breast cancer cohort.

Results: We identified signature genes and recurrence events which are predictive of recurrence free survival. The model was able to find unique genes that can differentiate between the grades. The identified genes for grade 1 are NEUROD2, IFNA14, SMCP, A1CF, CNTNAP1, APBB3, and G6PC2. For Grade 2, we identified CYP11B1, ARHGEF38, CHRNB3, and NTNG1. Whereas, grade 3 included NTNG1, ALS2CL, A1CF, and P2RY4 genes. We also found 2 genes A1CF and NTNG1 partcipating in grade 1& 3 and grade 2&3, respectively.

Discussion: The recurrence free survival model developed might aid in effective classification of tumor grades and recommending optimal therapy for the patients.

Introduction

Breast cancer is the most common type of cancer diagnosed in women in the United States[1]. The average 5-year and 10-year survival rate for people with breast cancer are 90% and 83% respectively[2]. An aggressive breast cancer can spread to lymph nodes, liver, lungs, bones, and even to the brain. Due to the aggressive nature of breast cancer cells, they are given a grade when they are removed from the breast and checked under a microscope for histopathology report[1].

The grade of a tumor is based on how much the cancer cells look like normal cells. Grade 1 cells are slower-growing, and look more like normal breast tissue. Grade 2 cells are growing at a faster speed and look like cells somewhere between grades 1 and 3. Grade 3 are cancer cells that look very different from normal cells and will probably grow and spread faster. The main drawback of current targeted cancer therapy is that they lack signature genes that can differentiate the grades, based on gene expression alone. This also limits personalized tailoring of targeted therapies to individual patients, and increases the chances of cancer recurrence[2].

Cancer recurrence is a phenomenon in which cancer returns after a period of remmission[3]. Recurrence usually occur in spite of the best therapies because some cells from cancer remain dormant and they decrease the survival chances of the patient, as recurrent cancers are more aggressive than the original. Therefore, identifying unique signature gene expression patterns, that increase the likelihood of breast cancer, for each grade from 1-3 can help clinicians to better strategise treatment protocol to prevent chances of recurrene. This may also help to better understand how likely it is that the treatment will be successful.

A number of inherited mutated genes that can increase the likelihood of breast cancer have been identified before, although they are not grade specific[4]. The most well-known are breast cancer gene 1 (BRCA1) and breast cancer gene 2 (BRCA2), both of which significantly increase the risk of both breast and ovarian cancer. Therefore, we worked on developing a multivariate statistical approach that prioratize cancer genes using information gain(I.G) and used structural equation modeling(SEM) to predict

signature genes which are most predictive of recurrence free survival in breast cancer patients.

Information Gain (IG) is a classical approach commonly used in computer science to determine which attribute or feature in a given data set are most useful for discriminating between some classes to be learned[5]. In our case, the attributes are genes and the class is 5-years recurrence. The use of IG for gene ranking in our study is achieved by calculating the information gain for the expression level of each gene. Genes are ranked high if their expression is information-rich in terms of recurrence, and are ranked low if their expression level is not informative for recurrence. It should be noted that this approach ignores other genes and how a gene is associated with other genes, but only considers the association of each gene's expression phenotypes.

Structural equation modeling is a multivariate statistical analysis technique that is used to analyze structural relationships[6]. This technique is the combination of factor analysis and multiple regression analysis, and it is used to analyze the structural relationship between measured variables and latent constructs. This method is preferred by the researchers because it estimates the multiple and interrelated dependence in a single analysis. The measurement model is the part which relates measured variables to latent variables. The structural model is the part that relates latent variables to one another. Statistically, the model is evaluated by comparing two variance/covariance matrices. SEM can take the best informative genes that can be identified using Information gain(I.G) and find the structural relationship between the vairables.

We aim to use these method to identify candidate genes which may be central to recurrence free survival. We anticipate that a ranked list of most informative genes would help us to focus our analysis on specific genes in the signaling pathways of the cell, which would help reduce the noise introduced by less important genes.

Materials and Methods

Microarray Data

The normalized gene expression data for 1,809 breast cancer patients were downloaded from KMplotter website[7]. More information about the data can be found at[7]. The patient data was divided into two groups based on the criteria of 'no recurrence' for more than 5 years and 10 years. The 5 year group contained 1519 patients. The data contained expression status of 22,222 probes. The probes were converted to their 11,000 gene symbols using affymetrix mapping. Further, the data was cleaned to remove the patients that doesn't have grade information. After missing value imputation, we selected 703 patients grouped into grade 1, grade 2 and grade 3. There were 137 genes in Grade 1, 333 genes in Grade 2 and 233 genes in Grade 3 respectively.

Gene ranking using Information gain

We used a classicial information gain(I.G) calculation method using R programming to find the importance of genes in our data. We applied the method to a compendium of breast cancer micro-array data. We used 5 years recurrence as the patients' class. The genes were ranked based on its information gain value and the top 23 genes were selected for further analysis.

Structural equation modeling (SEM)

Our SEM building consisted of 25 independent variables including 23 genes we selected using gene ranking, recurence free survival event (RFS_event) and death event. Whereas the dependent variable was recurrence free survival time (RFS_time) The SEM model was created for each grade separately, starting off with the same initial set of genes (**Figure 1**).

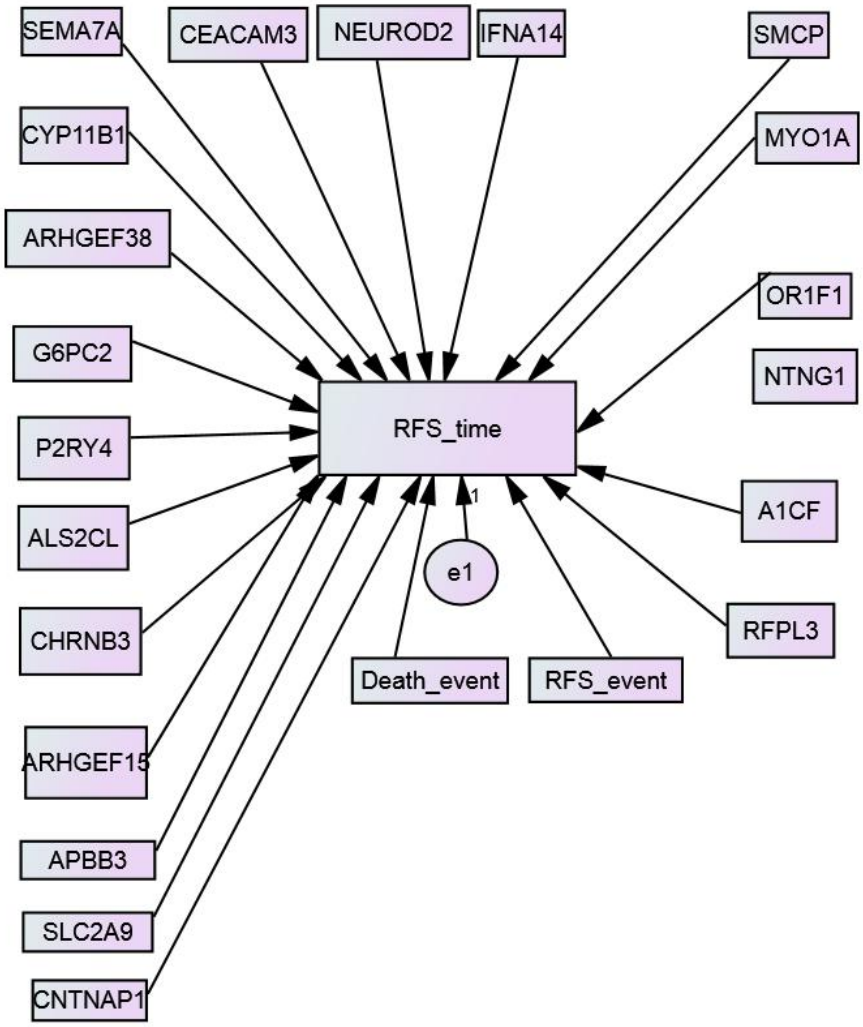

Figure 1. Initital structural equation diagram of the breast cancer gene expression model made using AMOS. There are 23 genes, death event and RFS event for all the models except for grade 2. There was not enough death event information available for grade 2 so it has been discarded from the analysis.

The goodness of fit of the model was estimated using chi-square test, p-value significance of the independent variables ($p<0.05$), CFI and RMSE values. At each step modification indices were used to find new structural relationship between the endogenous variables and model was refined and improved. The steps were repeated untill the best goodness of fit was obtained with all the independent variables showing significance and no more modification indices are suggested by AMOS. The model with the least chi-square value, highest CFI and lowest RMSE was selected as the final model. A simplified flowchart of the procedure can be seen in **Figure 2**.

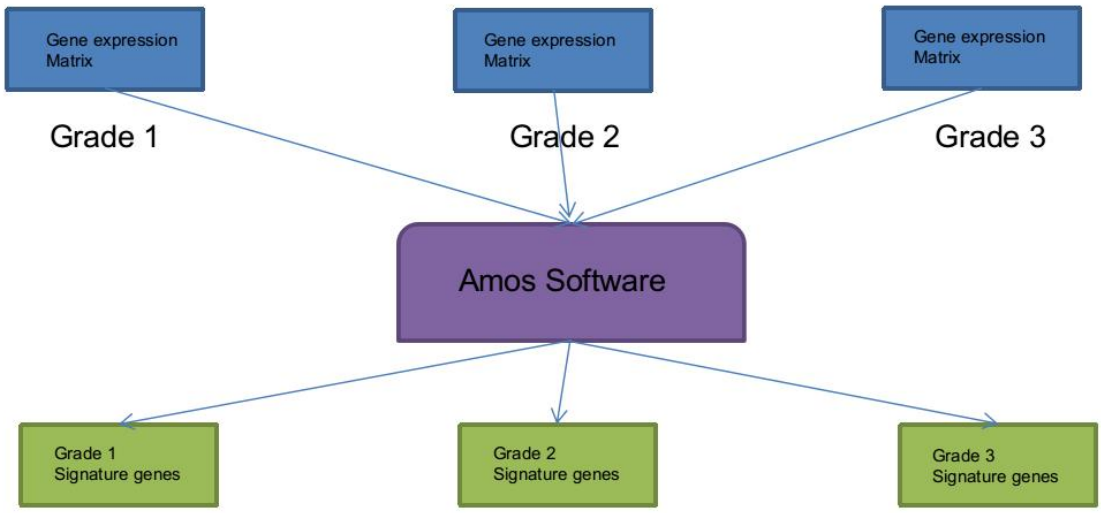

Figure 2. Flowchart of structural equation modeling for the patients with grades 1-3. Each grade data were fed into the AMOS software and an optimal model was identified through refinements. The best model with its signature genes were used for downstream analysis.

Results

Exploratory data analysis

Information gain was calculated to rank the genes of our study that are predictive of recurrence free survival time. Genes with an expression level from 0 to 300 was found to have the maximum frequency. The distribution of the expression of the genes are shown in **Figure 4 a**. The top 23 genes from this study were selected for SEM modeling.

The correlation matrix for all the variables has been calculated and shown in **Figure 4 c**. Hierarchical clustering was used to cluster the genes and the genes found to be grouped into three significant clusters as shown in **Figure 4 b**.

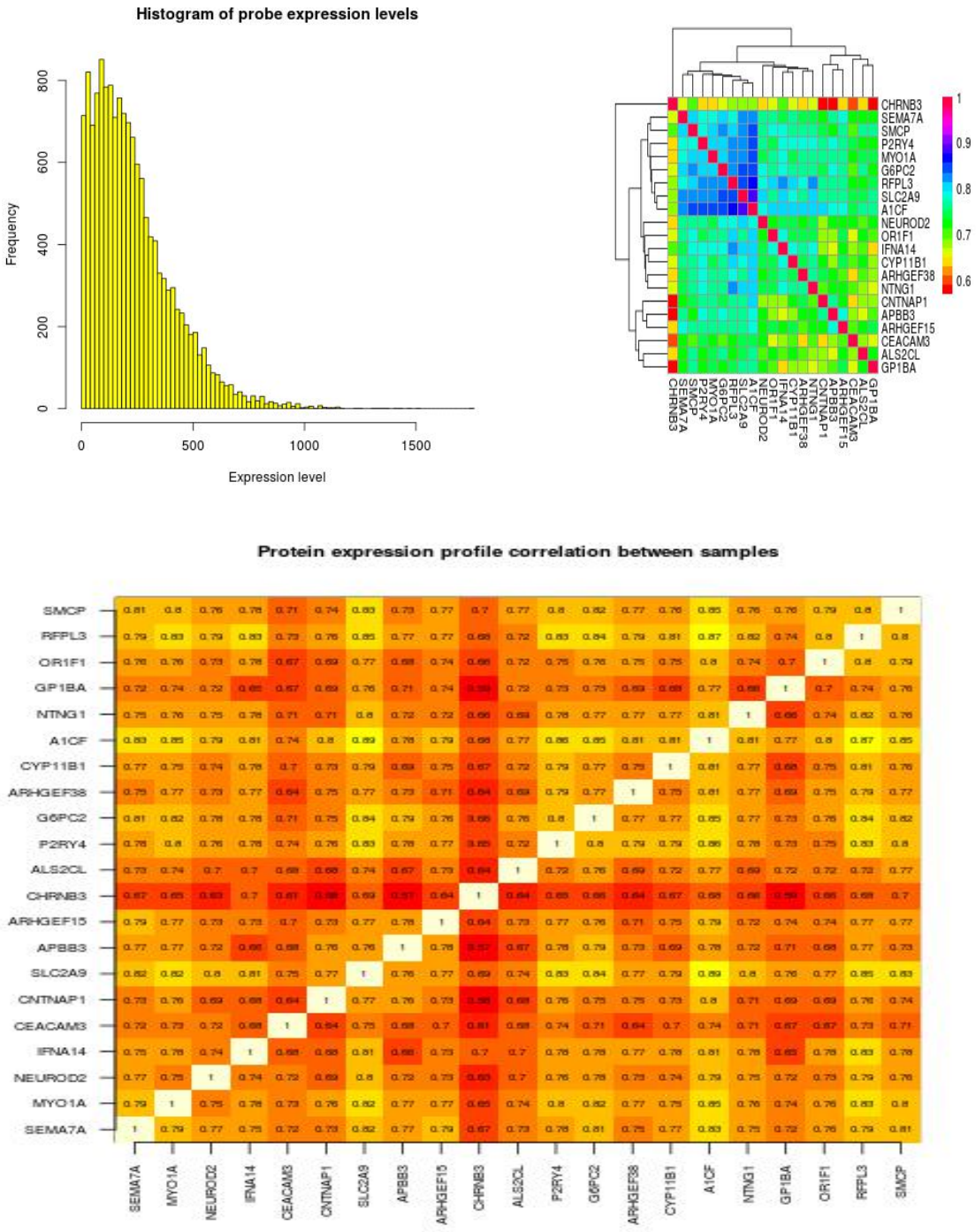

Figure 4 a) Gene expression distribution of top 23 genes selected for our study. Frequency disribution was maximum for genes with expression level beteen 0 and 300. **b)** Hierachical clustering showing the clustering of genes into three groups **c)** Correlation matrix of the top 23 genes selected for our study.

Structural equation modeling

Structural equation modeling was performed for breast cancer patients with grade1, grade 2 and grade 3. Incremental fit measures, including the NFI and CFI were used to assess the goodness of fit of the model. NFI and CFI determine how well a model fits compared with a baseline model. The estimates of model fit were calculated for each grade seperately.

The proposed SEM for each grade in this study has been shown in Figure 5, 7 and 9. The pathways from independent variables to RFS_time were all significant ($p < 0.05$). Table 1, 2 and 3 summarizes the results of estimates of regression weights of the proposed model.

Grade 1 (**Figure 5**) found to have $\chi 2 = 400.7$, Probability level = 0.0, NFI = 0.65, CFI = 0.66, RMSEA = 0.36, standardized RMR = 0.40. The model has 9 independent variables. We used the genes from this list for survival analysis to predict if high and low gene expression alone can differentiate patients and predict survival. The results are shown in **Figure 6.**

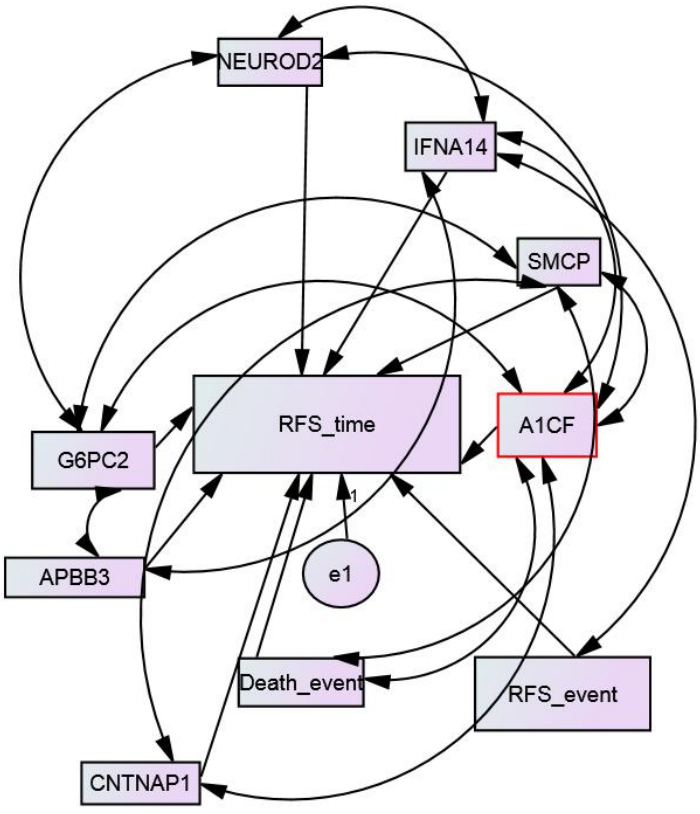

Figure 5. Structural equation model diagram for grade 1 breast cancer gene expression data. A total of 9 genes, death event and RFS event found to be predictive of recurrence free survival time. ($\chi 2$ = 400.7, Probability level = 0.0, NFI = 0.65, CFI = 0.66, RMSEA = 0.36, standardized RMR = 0.40)

Independent variable		Dependent variable	Estimate	S.E.	C.R.	P
RFS_time	<---	NEUROD2	.012	.004	3.239	.001
RFS_time	<---	IFNA14	-.009	.004	-2.523	.012
RFS_time	<---	SMCP	-.120	.035	-3.421	***
RFS_time	<---	A1CF	-.005	.002	-2.261	.024
RFS_time	<---	CNTNAP1	.005	.002	3.060	.002
RFS_time	<---	APBB3	-.003	.001	-2.832	.005
RFS_time	<---	G6PC2	.009	.004	2.543	.011
RFS_time	<---	Death_event	.000	.000	-2.311	.021
RFS_time	<---	RFS_event	-4.455	.510	-8.739	***

Table 1 : The results summarized the estimates of regression weights of the proposed model for grade 1.

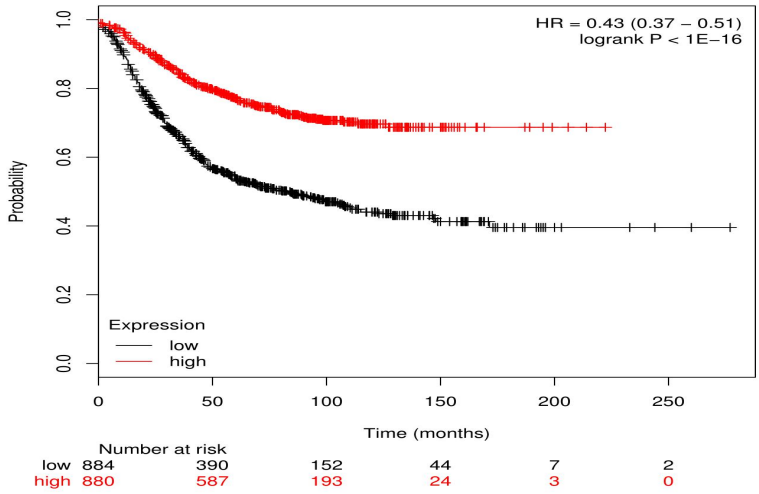

Figure 6. Survival analysis of genes for grade 2. Logrank p value found to be $1*10^{-16}$.

Grade 2 (**Figure 7**) found to have $\chi 2$ = 11.14, Probability level = .03, NFI = 0.99, CFI = 0.99, RMSE = 0.07**,** standardized RMR = 0.08. The model has 5 independent variables. We used the genes from this list for survival analysis to predict if high and low gene expression alone can differentiate patients and predict survival. The results are shown in **Figure 8.**

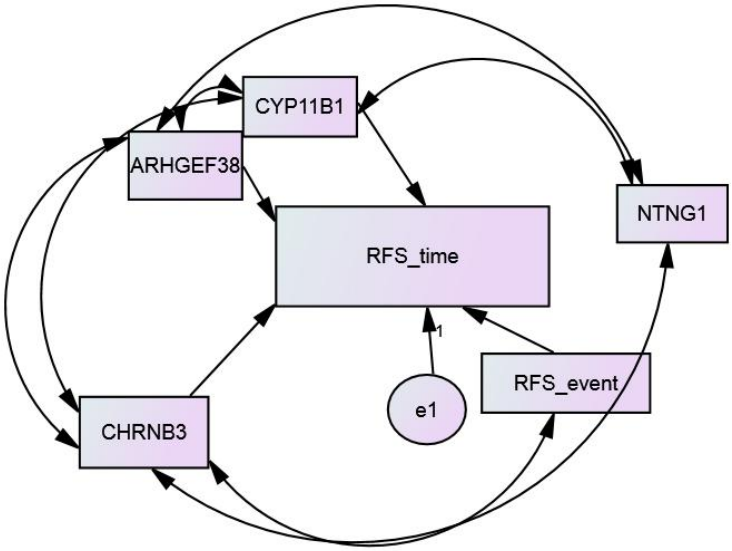

Figure 7. Structural equation model diagram for grade 2 breast cancer gene expression data. A total of 4 genes, and RFS event found to be predictive of recurrence free survival time. ($\chi 2$ = 11.14, Probability level = .03, NFI = 0.99, CFI = 0.99, RMSE = 0.07**,** standardized RMR = 0.08)

Independent variables		Dependent variables	Estimate	S.E.	C.R.	P
RFS_time	<---	ARHGEF38	-.004	.002	-2.042	.041
RFS_time	<---	CHRNB3	.016	.007	2.206	.027
RFS_time	<---	RFS_event	-5.551	.344	-16.141	***
RFS_time	<---	CYP11B1	-.007	.004	-1.868	.062

Table 2 : The results summarized the estimates of regression weights of the proposed model for grade 2.

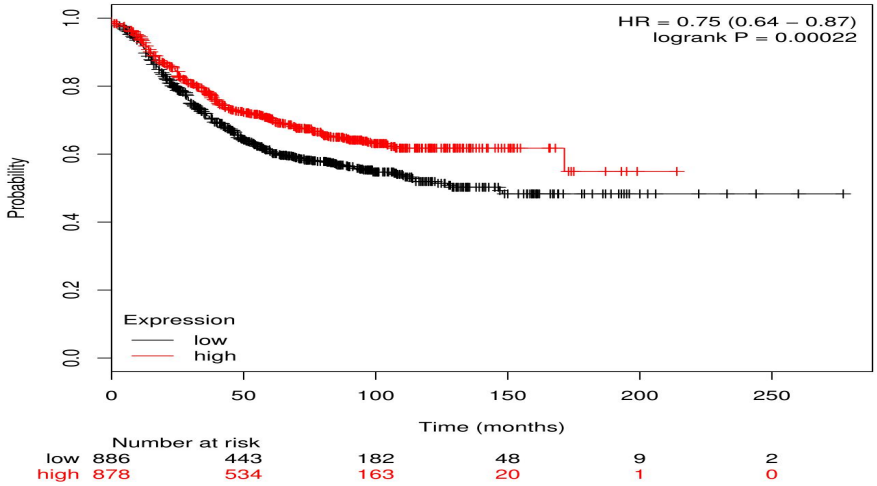

Figure 8. Survival analysis of genes for grade 3. Logrank p value found to be $2*10^{-4.}$

Grade 3 (**Figure 9**) found to have $\chi 2 = 30.46$, Probability level = 0.0, NFI = 0.97, CFI = 0.98 , RMSE = 0.10, standardized RMR = 0.12. The model has 6 independent variables. We used the genes from this list for survival analysis to predict if high and low gene expression alone can differentiate patients and predict survival. The results are shown in **Figure 10.**

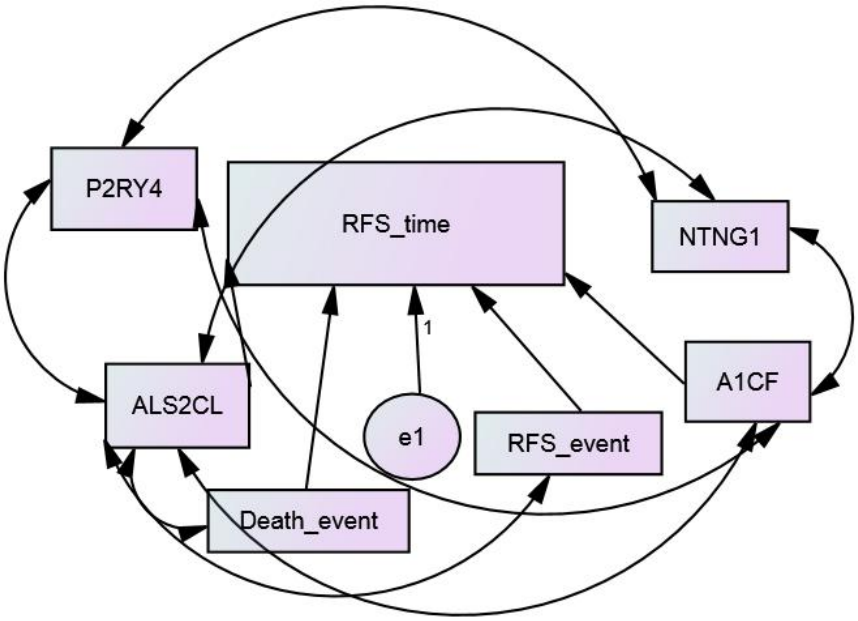

Figure 9. Structural equation model diagram for grade 1 breast cancer gene expression data. A total of 4 genes, death event and RFS event found to be predictive of recurrence free survival time. ($\chi 2 = 30.46$, Probability level = 0.0, NFI = 0.97, CFI = 0.98 , RMSE = 0.10, standardized RMR = 0.12)

Independent variables		Dependent variables	Estimate	S.E.	C.R.	P	Label
RFS_time	<---	A1CF	-.006	.002	-4.077	***	
RFS_time	<---	ALS2CL	.003	.002	1.749	.080	
RFS_time	<---	RFS_event	-8.012	.376	-21.285	***	
RFS_time	<---	Death_event	.000	.000	-2.154	.031	

Table 2 : The results summarized the estimates of regression weights of the proposed model for grade 3.

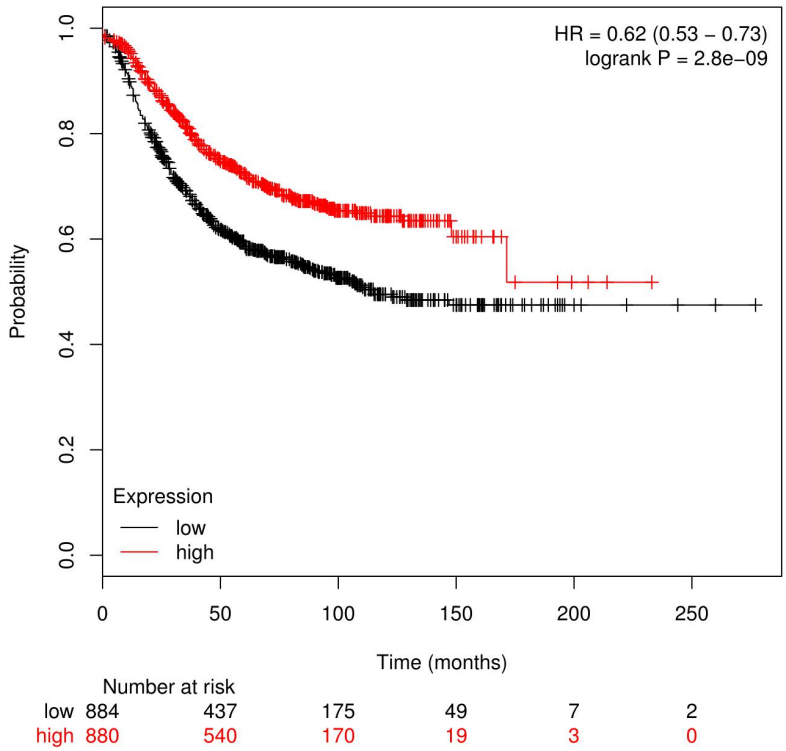

Figure 10 Survival analysis of genes for grade 3. Logrank p value found to be $2.8*10^{-9}$.

Discussion

Tumor grade is the description of a tumor based on how abnormal the tumor cells and the tumor tissue look under a microscope. It is an indicator of how quickly a tumor is likely to grow and spread. We used information gain and structural equation modeling to prioratize breast cancer genes & survival events and create an optimum model that finds direct and indirect structural relations within the model.

Information gain was used to rank the genes and the top 23 genes were used for downstream structural relationship modeling using AMOS software. The patients were grouped based on their grades from grade1 to grade 3 and the underlying signature genes and survival event were identified using structural equation modeling.

The exciting fact about our model is that we were able to find unique genes for each grade. Grade 1 has NEUROD2, IFNA14, SMCP, A1CF, CNTNAP1, APBB3, and

G6PC2. Grade 2 included CYP11B1, ARHGEF38, CHRNB3, and NTNG1. Whereas, grade 3 contains NTNG1, ALS2CL, A1CF, and P2RY4.

We found 2 genes A1CF and NTNG1 partcipation in more than one grade SEM. A1CF was found in grade 1 and grade 3 whereas, NTNG1 was found in grade 2 and grade 3. NTN1 or netrin-1 is a gene that has role in cell adhesion, motility, proliferation, and differentiation. Netrin-1 and its receptors, deleted in colorectal cancer and uncoordinated-5 homolog, have been linked to apoptosis and angiogenesis. Since these properties are essential for tumor development, Netrin-1 and its receptors have been reported to promote tumorigenesis in many types of cancers. Whereas, A1CF is reported to be high in liver and renal cancer.

We believe that this is the first study that has used a large sample size of ~700 patients in breast cancer to study the differene in tumore growth rate at grade specific level. The genes identified could be used as a signature genes in patients to determine the exact grade and give optimal treatment. However, we would like to admit that the SEM is more complex and finding ways to transfer the model to biology and making relationship at the biological pathway level might be really useful for finding patients who would really benefit.

References

1. https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html accessed on Feb 6th, 2018.
2. https://www.cancer.net/cancer-types/breast-cancer/statistics accessed on Feb 6th, 2018.
3. Yeh, Albert C., and Sridhar Ramaswamy. "Mechanisms of cancer cell dormancy—another hallmark of cancer?." Cancer research 75.23 (2015): 5014-5022.
4. Petrucelli, Nancie, Mary B. Daly, and Tuya Pal. "BRCA1-and BRCA2-associated hereditary breast and ovarian cancer." (2016).
5. Kent, John T. "Information gain and a general measure of correlation." Biometrika 70.1 (1983): 163-173.
6. Byrne, Barbara M. Structural equation modeling with AMOS: Basic concepts, applications, and programming. Routledge, 2016.
7. http://kmplot.com/analysis/ accessed on December 13th, 2017.