International Journal of Mathematical Sciences in Medicine(IJMSM) Vol 1, No2, 2018. Page 23–32

Analysis Seers Breast Cancer Data Using Random Forest

Yong Xu[1] and Bong-jin Choi[2]

[1]Ferris State University, Marketing Department
Big Rapid, MI 49307, USA
[2]North Dakota State University
Fargo, ND 58108-6050, USA

ABSTRACT: The object of the present study is to conduct a machine learning analysis for breast cancer tumor size prediction for United States patients based on real uncensored data. Based on the result of this analysis, we will develop a statistical model with only important variables. We accomplish the objective by developing a high quality statistical model that identifies the significant attributable variables and interactions. We rank these contributing entities according to their percentage of the contribution to breast cancer tumor growth. We used this new machine learning way to help us update one of previous model which is developed through classical modeling method. This proposed statistical model can be used to conduct surface response analysis to identify the necessary restrictions on the significant attributable variables and their interactions to minimize the size of the breast tumor. One can also use the proposed model to make early prediction of tumor size based on the most attributable variables.

1 INTRODUCTION

Wikipedia [20] (the online encyclopedia) defines breast cancer as: a cancer that starts in the breast, usually in the inner lining of the milk ducts or lobules. There are different types of breast cancer, with different stages (spread), aggressiveness, and genetic makeup.

The proposed model that we are developing includes individual variables and interactions. The response variable we use to develop this statistical model is the tumor size at diagnosis for breast cancer patients. We have included 26 possible attributable variables for breast cancer, namely, X1, X2,..., X26. For instance, X1 stands for patient ID and X2 stands for the patients age (in years) when he/she got the diagnosis. The full list of attributable variables is in Table 1.1, below. In this research, we would like to find the relation between the tumor size and all other attributable variables. We cannot use survival time to predict the tumor size since death time happens after the tumor is detected. Therefore, we exclude the variable survival time(x25) and the censoring indicator function vss (x26) in the first part of study. Thus, we have only 24 variables left to construct our statistical model. Due to similar reasons, we have to exclude variable x23, COD (Cause of Death) since we do not possess this information when the diagnosis of tumor happens. Therefore we will use only 23 variables for our study.

Findings from multiple studies have demonstrated that one result of participation in quality childcare programs is improvement in children's receptive and expressive language skills (Goelman & Pence, 1987; Howes, 1997; McCartney, 1984). It is therefore of great importance that childcare providers are knowledgeable about and have the skills to stimulate and promote language development in their centers. According to Mroz and Hall (2003), however, only 16% of early childhood providers believe that they have received adequate training in the area of child language to provide quality services in this developmental domain. Forty-six out of fifty childcare providers who were interviewed expressed a need for training specifically in the area of speech and language. The providers noted that ideal content for this training would consist of identification of children with speech and language difficulties within the context of a basic introduction to speech and language development (Mroz & Hall, 2003).

In the present analysis, we used real data from the Surveillance Epidemiology and End Results (SEER) Program. SEER collects and compiles information on incidence, survival, and prevalence from specific geographic areas representing about 26 percent of the U.S. population plus cancer mortality for the entire U.S. [19].

The proposed statistical model is useful in predicting the tumor size given data for the attributable variables. It is statistically evaluated using R square, R square adjusted, the PRESS statistic and several types of residual analyses. Finally, its usefulness is illustrated by utilizing different combinations of the attributable variables. We use one popular machine

Received October 10, 2018 IFNA Publishers, Inc.

learning method: random forest to help us select the attributable variables and then construct the statistical model based on that. In addition, the attributable variables are ranked according to their contributions to accurately estimate a patients tumor size.

2 HISTORICAL REVIEW

Survival analysis is used more and more in the areas such public health, medicine, medical research, financial research, etc. Many researchers have made signficiant contributions to this subject. For example, C. A. McGilchrist, [15], [16] discussed the regression with frailty in survival analysis. D. R. Cox, [4] introduced the Cox proportional hazards (PH) model for survival data. E. L. Kaplan and P. Meier, [10] constructed Kaplan Meier empirical type of survival model. E. A. Gehan developed the generalized Wilcoxon test and this test is more powerful than the Cox proportional hazards test when the proportional hazard assumption is violated early on. K. Liu and C. P. Tsokos, [11], [12], [13] utilized kernel density methods in reliability analysis. N. Mantel and W. Haenszel, [14] proposed the Mantel Haenszel test for survival analysis. P. Qiu and C. P. Tsokos studied extensively the accelerated life testing model. Y. Xu and C. P. Tsokos, [23] probabilistically discussed and evaluated several most commonly used survival analysis models. Y. Xu and C. P. Tsokos, [21] use classical statistical modeling method construct a statistical model and ranked the contribution of the important variables. After we have the result from the machine learning, we will compare our result with this paper [21] and proposed a modified version of the statistical model.

Some classical historical research papers can be found in [5], [7], [10] and [17]. Other important and recent references for the readers who will have an interest in survival analysis may be found in [2],[3],[6], [8], [18] and [22].

Table 1.1 List of attributable variables

Name	Full name of variables	Short form
X1	Patient ID	id
X2	Age at diagnosis	age
X3	Year of Birth	birthy
X4	Birth Place	birthp
X5	Sequence Number Central	snc
X6	Month of diagnosis	month
X7	Year of diagnosis	year
X8	Primary Site	ps
X9	Laterality	la
X10	Histologic Type ICDO3	ht
X11	Behavior Code ICDO3	bc
X12	Type of Reporting Source	trs
X13	RXSumm SurgPrimSite	rxps
X14	RXSumm Radiation	rxr
X15	RXSumm RadtoCNS	rxcns
X16	Age Recode Year olds	ager
X17	Site Recode	sr
X18	CSS chema	css
X19	AJCC stage 3rd edition	ajcc
X20	First malignant primary indicator	findi
X21	State-county recode	scr
X22	Race	race
X23	Sex	sex

3 DEVELOPMENT OF THE NONLINEAR STATISTICAL MODELS

We randomly extract 155 uncensored breast cancer patients information from SEER data base. The data was obtained from 2000 to 2006. We want to develop a statistical model with full information instead of censored. Therefore, we will use the 155 uncensored patients information to construct our statistical model.

We proceed to develop a statistical model taking into consideration the twenty three attributable variables listed in Table 3.1. The form of the statistical model is given by tumor size as a function of (x1, x2, , x23). Note that some of the variables values are obtained after the tumor size is recorded. In our analysis all the patients in the data base have breast cancer. We utilize the values of the tumor size once the patient has gone through a diagnostic process. Thus, the general statistical form of the proposed model with all possible attributable variables and interactions will be of the form in equation 3.1.

$$TS = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + \cdots + \alpha_i A_i + \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_j B_j + \epsilon \tag{3.1}$$

Here, TS stands for tumor size, the α and β are the coefficients, and A is the first order term of the attributable variables and B are the possible interactions and higher order terms. The object is to develop the most representative estimate of the above model based on available data.

Here variable x 19 AJCC refers to the 3rd edition of the AJCC cancer staging manual from American joint committee on cancer and we use the same standard for all patients that we have.

One of the basic underlying assumptions in formulating an estimate of the above statistical model is that the response variable should be Gaussian distributed. Unfortunately, in the present form that is not the case. This fact is clearly demonstrated by the QQ plot shown by Figure 3.1, below.

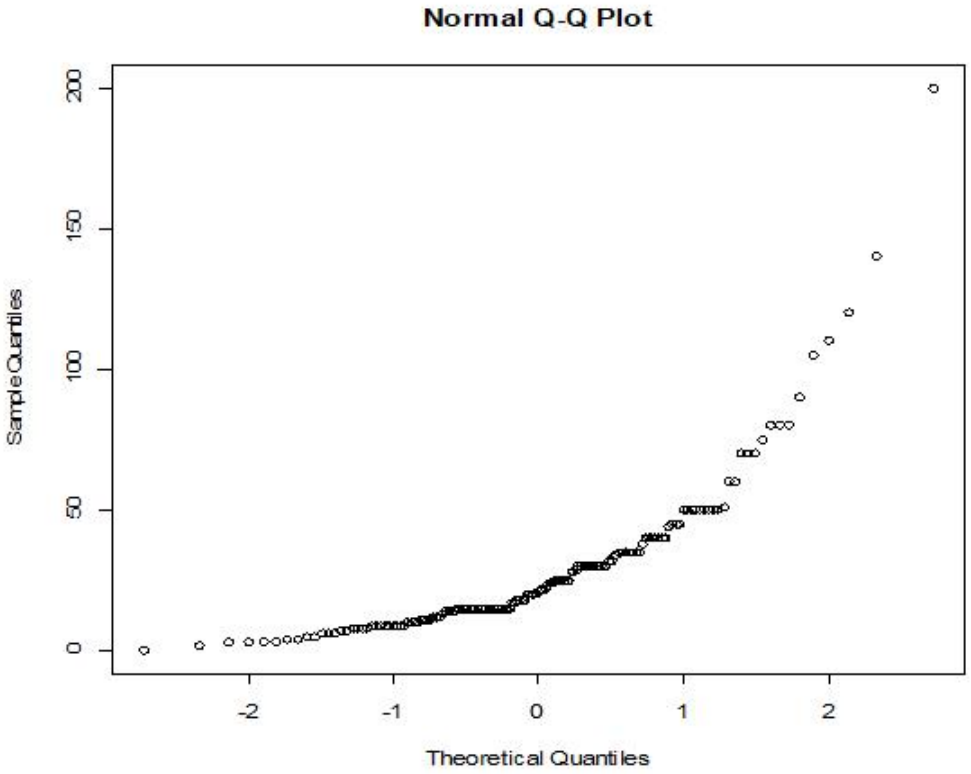

Figure 3.1 QQ Plot for Testing Normality for Original Data

Furthermore, the Shapiro-Wilk normality test with the necessary calculation of the test statistic W = 0.7437 and p-value = 3.787e-15 is additional evidence that the tumor size does not follow normal probability distribution. We proceed in utilizing the Box-Cox transformation to the tumor size to determine if such a filter will modify the given data to follow the normal distribution so that we can proceed to formulate the proposed statistical model. After we applied the Box-Cox transformation, we can use the transformed tumor size as our new response variable. Thus, we can proceed to estimate the coefficients of the attributable variables for the filtered transformed tumor size data to obtain the coefficient of all possible interactions and at the same time determine the significant contributions of both attributable variables and interactions.

We begin with the previously defined twenty-three attributable variables x1, x2, . . . , x23 and more than two hundred 2nd and 3rd order interaction between each pair, and we did not consider any 4th and higher order interactions due to the initial research from previous study [21]. Since we already have a lot more terms than patients in the data itself, we utilized sub modeling skills and two way selection procedures to construct our model.

We divided the tumor size into two different categories: small for tumor size less than 20 mm and large if the tumor size is greater or equal to 20mm. We further divided our data into two sub data set. The first data set is randomly selected from the original data set and it has about 60% of the original data set. We call it as train data set and we will use it to build our random forest models. The second data set is the remaining data set from the original and we call it test data set. We will use test data set to run our random forest model and check the goodness of the prediction. From table 3.1, below, we can see the estimation error rate is about 20% and Out of Bag (OOB) estimate of error rate is 21.51%. It shows that our random forest estimation has a high accurate rate on discrimination.

Table 3.1 Confusion Matrix

Size	Large	Small	Class.error
Large	37	9	.1957
Small	11	36	.2340

To develop the models, initially we start building our model with random forest process with up to 500 trees. From Figure 3.2, below, we can see with 200 trees we have had a stable error rate. Therefore, we will use 200 trees to develop our models. The green dot plot is for big tumor group. The dash red plot is for small tumor group and black solid plot is for combined group.

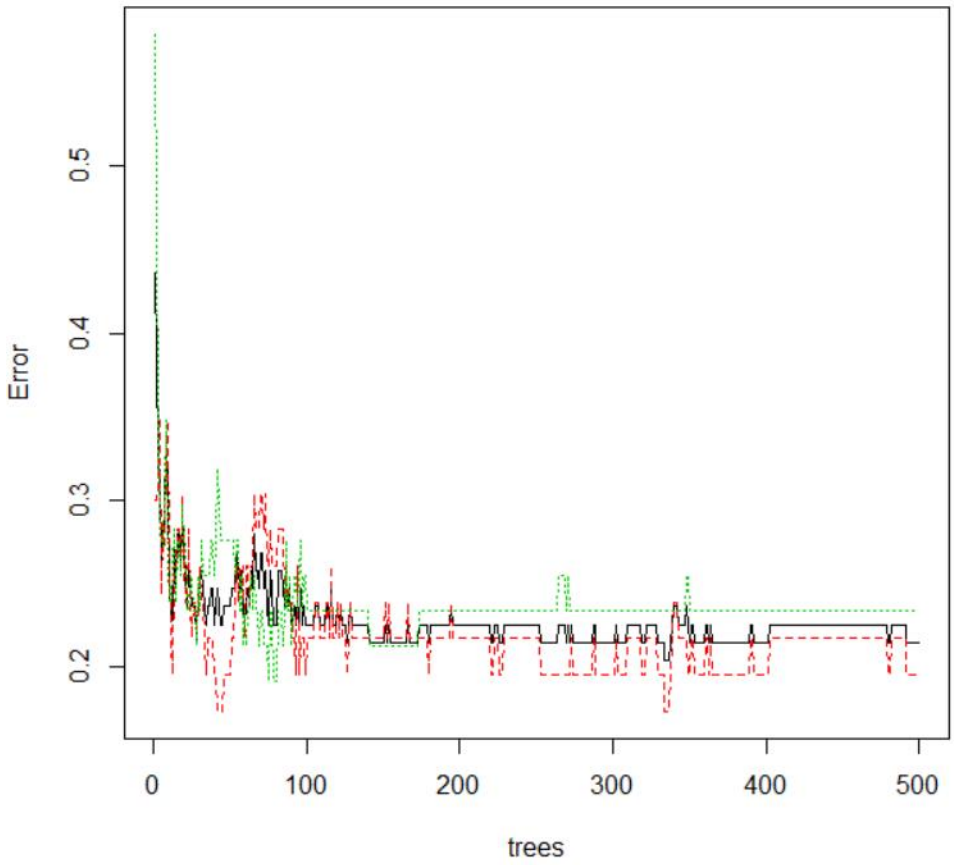

Figure 3.2 Breast Cancer Data random forest tree plot

From Figure 3.3, blow, we can see as the mean square analysis show, the number of predictors should not more than 10 as the error will not further decrease when we have more variables enter the random forest analysis. Therefore we will keep our number of variables as 10 for each trail.

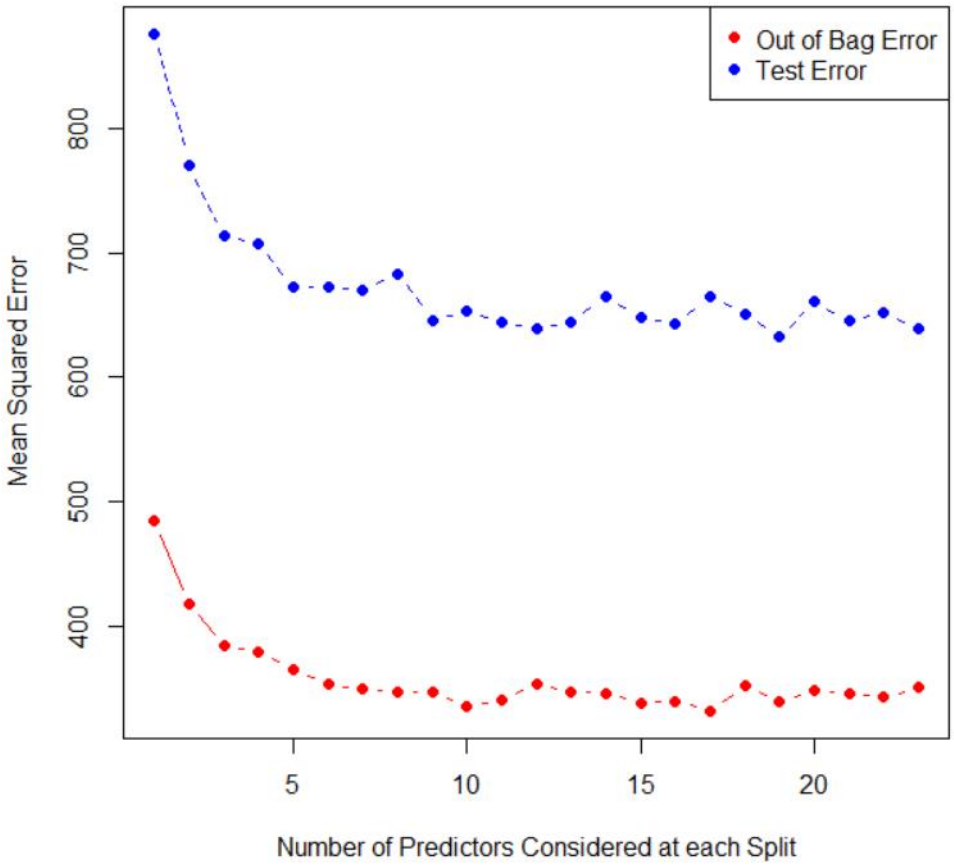

Figure 3.3 Number of variables at each split in random forest analysis

Figure 3.4, below, give us the important attributable variables that are selected by random forest process. Based on both methods, we decide to pick the following 10 attributable random variables to construct our model: AJCC(X19), YEAR(X7), AGE(X2), BIRTHY(X3), MONTH(X6), SNC(X5), ID(X1), SEX(X23), RXR(X14) and RXPS(X13). Out of these random variables, AGE, BIRTHY are related and they refer to the current age of the patients. Therefore, we will combine these effects and use only one variable AGE to represent the patients actual age. With similar reason, variables Month and Year are both used to represent the age when patients make the initial diagnosis. This age is quite different from their current age. To reduce the dependence of the variables, we will only keep variable Year alone. SEX and ID are nuisance factors. So the final variables that we want to keep are AJCC(X19), YEAR(X7), AGE(X2), SNC(X5), RXR(X14) and RXPS(X13). Compared this result with research [21], we can see the previous study suggest RXR(X14), COD(X23), RXPS(X13), SNC(X5) and AJCC(X19) as most significant variables which is almost identical with our finding here. The only new variables that we find are YEAR and AGE.

We first construct an initial model with no interactions and just first order terms as the following: AJCC(X19), YEAR(X7), AGE(X2), SNC(X5), RXR(X14) and RXPS(X13). After the initial model that we constructed, we display the tree with our model. From Figure 3.5, below, we can observe the error rate indeed converge toward low rate when the numbers of trees approach 200 as what we expected before.

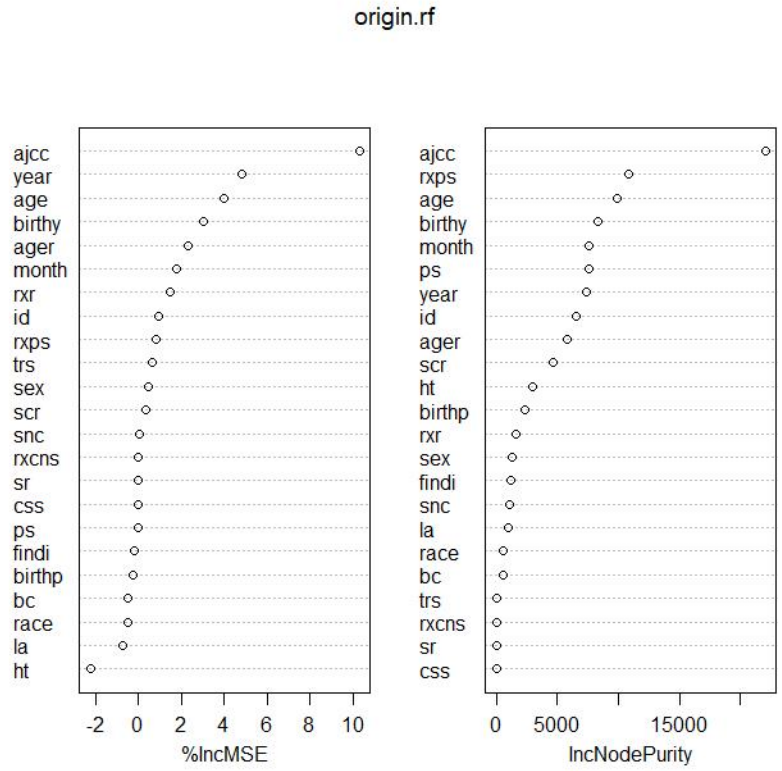

Figure 3.4 Important attributable random variables from random forest analysis

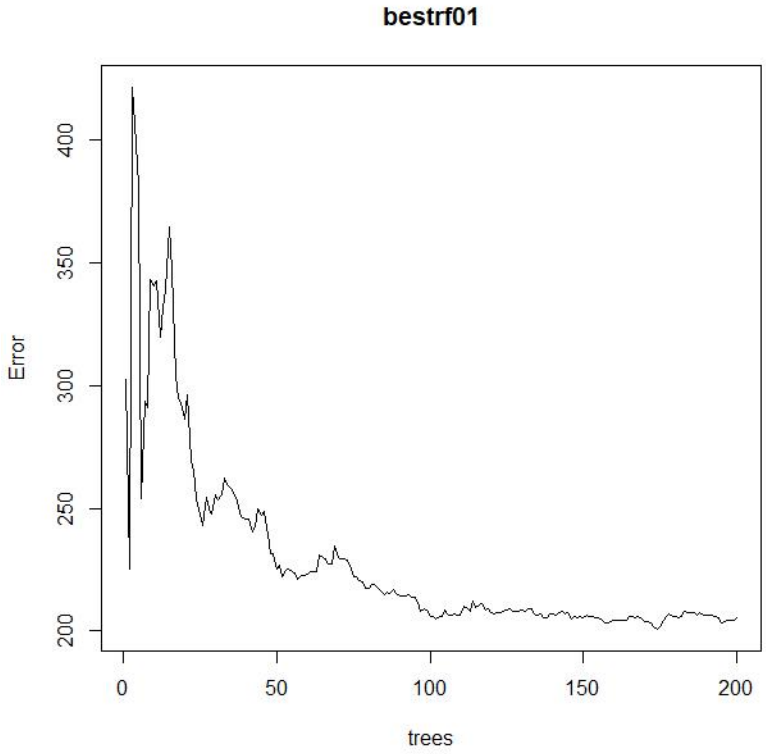

Figure 3.5 Error rate vs number of trees from random forest analysis

There are several missing values in the variable AJCC. We use the mean of the rest of the data value in the variable AJCC to replace the NA value in order to perform prediction of the model. We add interaction term into our model and redo the random forest analysis. HT(X10) is not considered as significant variable in our initial study. However, after our random forest analysis for interactions, we found the same three most significant interaction terms as our previous

research [21]. Therefore we will include HT in our model even with none significant first order term but as significant higher order term. After we figure out all the single term and high order term, we perform a classical regression method to identify all the estimates of the coefficients of our random variables. Thus, the results of estimation of equation 3.1 are given by equation 3.2 as follows

$$\ln(\widehat{TS}^{0.2659}) = 88.51 - 2.16 * 10^{-3}X_2 - 2.85 * 10^{-2}X_5 + 4.57 * 10^{-2}X_7 - 2.61 * 10^{-3}X_{10} \\ +4.1 * 10^{-3}X_{13} - 3.24X_{14} + 3.1 * 10^{-3}X_{19} + .24X_{14} * X_{19} + 1.86 * 10^{-4}X_5X_{10}X_{14} - 0.119X_5X_{14}X_{19} \tag{3.2}$$

Where X2 is AGE, X7 is YEAR, X5 is SNC, X10 is HT, X13 is RXPS, X14 is RXR and X19 is AJCC.

Here we keep HT even it is not significant individually. But we still keep it because HT is included in interaction terms. We will utilize the initial transformation that we used to transform the response data to get the result in equation 3.3 by taking the exponential on both sides of the 3.2 and then take 3.76s power on both sides of the equation 3.2.

$$\widehat{TS} = exp(88.51 - 2.16 * 10^{-3}X_2 - 2.85 * 10^{-2}X_5 + 4.57 * 10^{-2}X_7 - 2.61 * 10^{-3}X_{10} \\ +4.1 * 10^{-3}X_{13} - 3.24X_{14} + 3.1 * 10^{-3}X_{19} + .24X_{14} * X_{19} + 1.86 * 10^{-4}X_5X_{10}X_{14} - 0.119X_5X_{14}X_{19})^{3.76} \tag{3.3}$$

For our final model, the R squared is .894 and the previous final models R square is 0.889. The R squared adjusted of our final model is 0.887 and the previous final models R square adjusted is 0.881. We can see our model is comparable with the previous research results [21]. Both R squared value and R squared adjusted value are high (close to 90%) and these two values are very close to each other. This shows our models R squared increase is not due to the increase of the parameters estimates, but rather the good quality of the proposed model to predict tumor size given values of the identified attributable variables [1]. Secondly, the prediction of residual error sum of squares (PRESS) statistics evaluate how good the estimation is when we remove one data point at a time. We will choose the model that has the smallest PRESS statistics. The PRESS statistics results support the fact that the proposed model is of high quality. Our current models PRESS value is 89.25479 while previous final models PRESS value is 96.73797 which show obvious improvement.

Furthermore, we compared the Akaike Information Criterion (AIC) of our proposed model and our previous final model. Our current models AIC is 39.77 and our previous final models AIC is 46.35 which show significant improvement as well.

In Table 3.4 we have listed all the important attributable variables and interactions. We notice some variables are not significant by themselves but we still include them. For example, X5 and X10 are not significant by themselves, only in combination with the others. We summarize the attributable variables individually and interactively in the following schematic network as showed in Figure 3.6.

Table 3.4 List of Attributable Variables

No	Individual variables	Name of individual variables
1	X2	AGE of patients
2	X5	Sequence Number Central
3	X7	Year of diagnosis
4	X13	RXSumm SurgPrimSite
5	X14	RXSumm Radiation
6	X19	AJCC stage3 rdedition
Interactions		
7	X14:X19	RXR ∩ AJCC
8	X5:X10:X14	SNC ∩ HT ∩ RXR
9	X5:X14:X19	SNC ∩ RXR ∩ AJCC

We summarize the attributable variables individually and interactively in the following schematic network as showed in Figure 3.6.

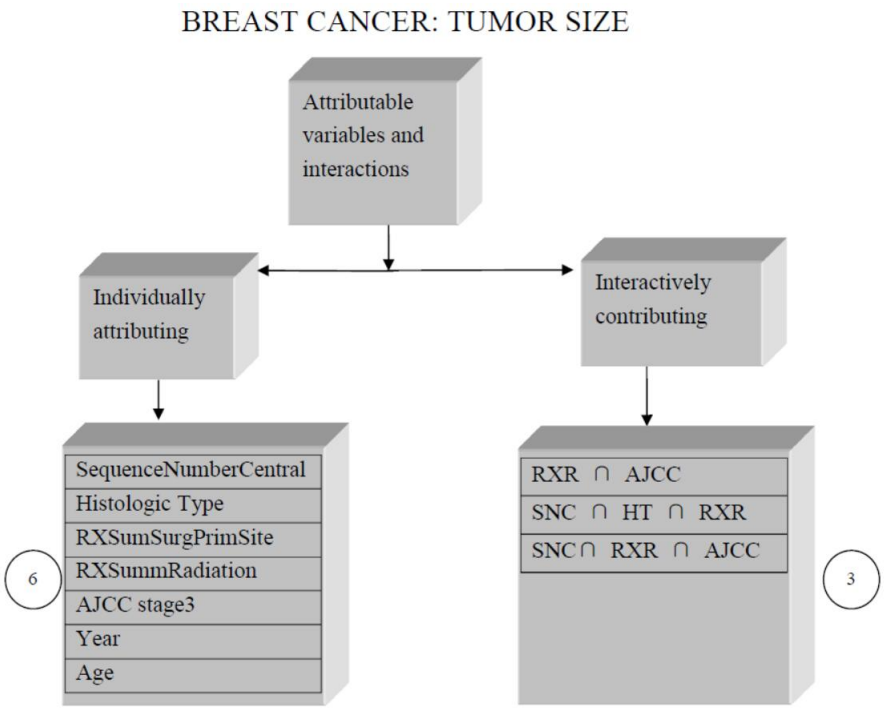

Figure 3.6 Breast Cancer Attributable Variable Diagram

In Table 3.5 are the ranks of the most top ten most significant attributable variables with respect to their contribution to estimating tumor size. The interaction among X5, X14 and X19 ranks as the top one. This is the interaction among the sequence number central(SNC), RXsumm Radiation(RXR) and AJCC stage3 rdedition(AJCC) [1].

Table 3.5 Rank of Variable According to Contributions

Rank	Variables
1	X5:X14:X19
2	X14:X19
3	X5:X10:X14
4	X19
5	X2
6	X7
7	X5
8	X13
9	X14
10	X10

4 USEFULNESS OF THE PROPOSED STATISTICAL MODEL

We can conclude from our extensive statistical analysis that there are only five significant attributable variables to the tumor size for breast cancer namely, YEAR(X7), AGE(X2), RXR(X14), RXPS(X13) and AJCC(X19). As for SNC(X5) and HT(X10), they themselves individually do not significantly contribute to the response variables; however, when they interact with other variables, they do significantly contribute to the response variable. Compare with previous study [21], we found two new attributable variables, namely, Age and Year. Furthermore, we also tested higher order interaction terms of the attributable variables and we found three interactions to significantly contribute to tumor size for breast cancer.

This model is useful for a number of reasons.

1. It can be used to identify the significant attributable variables which might contribute to the form of the tumor in breast.

2. It identifies the significant interactions of these attributable variables which might contribute to the interaction effects in breast cancer.

3. The most significant variables that contributions to the tumor size growth are ranked.

4. One can also use the proposed model to generate various scenarios of the tumor size as a function of different values of the subjective entities.

5. A confidence interval for the tumor size can be constructed with parametric analysis. By obtaining the % confidence limits for the response, we can describe how confident we are that our estimate is close to the actual tumor size.

6.The model as shown in equation (3.3) can be used to perform surface response analysis to place the restrictions on the significant attributable variables and interactions to minimize the breast cancer tumor size. We can also put restrictions on the variables to minimize the response of the tumor size using nonlinear control methods.

5 CONCLUSIONS & DISCUSSION

In the present study, we performed random forest analysis followed by a parametric analysis to estimate tumor size for breast cancer patients. The initial measurement of tumor size was collected from the SEER database. Those data do not follow normal probability distribution. Using the standard Box-Cox transformation, the SEER tumor size data became approximately normally distributed. We conducted machine learning method random forest analysis to select the most significant attributable variables. We developed a nonlinear statistical model based on these significant attributable variables. Through the process of developing the statistical model, we found only five variables, namely, YEAR(X7), AGE(X2), RXR(X14), RXPS(X13) and AJCC(X19) and three interactions that significantly contribute to the tumor size. The proposed statistical model was evaluated using the R-square, R-square adjusted, PRESS statistics and AIC methods, all of which support the high quality of the developed statistical model. We compared our model with previous research [21] and find significant improvement. This model can be used to obtain a good estimate of tumor size knowing the significantly attributable variables and interaction terms.

REFERENCES

[1] B. Abraham and J. Ledolter, *Introduction to regression modeling*, (2006)

[2] Bong-Jin Choi,"Statistical Analysis, Modeling, and Algorithms for Pharmaceutical and Cancer Systems" (2014). Graduate Theses and Dissertations. https://scholarcommons.usf.edu/etd/5200

[3] D. Collett, *Modeling survival data in medical research* , Chapman & Hall/CRC, (2003)

[4] D. R. Cox, *Regression models and life-tables (with discussion)*, Journal of the Royal Statistical Society Series B. (1972) 34: 187-220.

[5] A. W. Fyles, D. R. McCready, L. A Manchul., M. E. Trudeau, P. Merante, M. Pintile, L. M. Weir, and I. A. Olivotto, *Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer* , New England Journal of Medicine (2008), 351, 963-970.

[6] E. A. Gehan, *A generalized Wilcoxon test for comparing arbitrarily singly-censored samples*, Biometrika, (1965),52, 1 and 2, 203-223.

[7] D. P. Harrington and T.R. Fleming, *A Class of Rank Test Procedures for Censored Survival Data*, Biometrika, (1982), 69(3):553-566.

[8] N. A Ibrahim, A. Kudus, I.Daud, and M. R. Abu Bakar, *Decision tree for competing risks survival probability in breast cancer study*, International Journal of Biomedical Sciences Volume 3 Number 1, (2008).

[9] E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, (1958), 53:457-448.

[10] J. P. Klein, *Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm*, Biometrics, (1992), 48(3)795-806.

[11] K. Liu and C. P. Tsokos, *Nonparametric Density Estimation for the Sum of Two Independent RandomVariable*, Journal of Stochastic Analysis, (2000)

[12] K. Liu and C. P. Tsokos, *Nonparametric Reliability Modeling for Parallel Systems*, Journal of StochasticAnalysis, (1999).

[13] K. Liu and C. P. Tsokos, *Optimal Bandwidth Selection for a Nonparametric Estimate of the Cumulative Distribution Function*, International Journal of Applied Mathematics, (2002), Vol.10, No.1, pp.33-49.

[14] N. Mantel and W. Haenszel, *Statical aspects of the analysis of data from retrospective studies of disease*, Journal of the National cancer Institute, (1959), 22(4).

[15] C. A. McGilchrist and C.W. Aisbett, *Regression with Frailty in Survival Analysis*, Biometrics. (1991) 47(2):461-466.

[16] C. A. McGilchrist, *REML Estimation for Survival Models with Frailty*, Biometrics, (1993), 49(1):221-225.

[17] P. Qiu and C. P. Tsokos, *Accelerated Life-Testing Model Building with Box-Cox Transformation*, Sankhya, (2000), Vol. 62, Series A, Pt. 2, pp. 223-235.

[18] C. P. Tsokos, and Y. Xu, *Statistical Modeling of Breast Cancer using Differential Equations*, Istanbul University Journal of The School of Business Administration, (2011), Vol:40, No:1, 60-71.

[19] U.S. National institutes of Health, *http://seer.cancer.gov*

[20] Wikipedia *http://en.wikipedia.org/wiki/Breast_cancer*

[21] Y. Xu, J. Keper, and C. P. Tsokos, *Identify attributable variables and interactions in breast cancer*, Journal of applied sciences, (2011), 11 (6): 1033-1038.

[22] Y. Xu, and C. P.Tsokos, *Non-homogenous Poisson Process for Evaluating Stage I & II Ductal Breast Cancer Treatment*, Journal of Modern Applied Statistical Methods, (2011), Vol. 10, No.2, 646-655.

[23] Y. Xu, and C. P. Tsokos, *Probabilistic Survival Analysis Methods using Simulation and Cancer Data*, Problems of Nonlinear Analysis In Engineering Systems, English/Russian, (2012), 1(37), v.18, 47-59.