Predictors for Breast Cancer Recurrence

ABSTRACT

Breast cancer is one of the most common cancers in women and its recurrence has been linked to a variety of biological factors. In this project, we built an optimal logistic regression model in R, based on AIC and BIC criteria with stepwise procedures, to predict the recurrence outcome of a breast cancer patient. Using real patient data from the 1988 Ljubljana Breast Cancer Data Set, our model accurately discriminates and predicts 77.26% of the cases when applying leave-one-out cross validation technique with a 50% probability as a cutoff for recurrence or no recurrence event.

Background and Significance

Breast cancer is one of the most common cancers in women and a leading cause of cancer death [1]. Even more, over the last decade breast cancer risk has increased in certain demographic groups. As such, there remains a need to remove the cancer early to reduce recurrence. Since recurrence within 5 years of diagnosis is correlated with chance of death, being able to understand and predict recurrence susceptibility is critical [2]. Biologically, certain factors have been linked to this recurrence. For example, more cancerous lymph nodes at time of tissue removal has been found to increase recurrence risk [3]. General breast cancer risk has also been found to increase up until menopause after which this increase slows down [4]. For treatment, particularly in individuals with recurrence, radiation may be used to shrink tumors and kill cancerous cells [5].

With these known factors, we want to both verify and better understand what might influence the probability of a patient's breast cancer recurrence. More specifically, our focus question is: what factors make those in remission more susceptible to a breast cancer recurrence event? We wanted to explore the possible influence of variables such as age, number of cancerous lymph nodes, menopause status, and radiation treatment on recurrence probability. Based on previous research, we expected increased risk with higher age and number of cancerous lymph nodes particularly in premenopausal patients who received radiation treatment.

Table 1. *Variable names and descriptions*

Name	Description
Age	age (in years at last birthday) at time of diagnosis
PostMeno	whether the patient is pre- or post-menopausal at time of diagnosis
TumorSize	the greatest diameter (mm) of the removed tumor
#InvNodes	number of axillary lymph nodes with visible metastic breast cancer at time of diagnosis
NodeCaps	whether the cancer metastasized to a lymph node or not
Deg-Malig$_i$	histological grade (range 1-3) of the removed tumor; e.g. Deg-Malig$_3$ indicates grade 3
Breast	left or right breast where tumor occurred
Quadrant	location of tumor within breast (upper left, upper right, central, lower left, or lower right)
Radiation	whether the patient received radiation therapy or not

Data

Our data are from the Ljubljana Breast Cancer Data Set available from the UCI Machine Learning Repository [6]. The data were collected from 286 patients at the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia in July 1988 [7]. All patients had undergone surgery to remove cancer tissue. Specifically, in the dataset are nine predictor variables and one response variable indicating if a patient had any recurrence event(s) within five years post-operation (Table 1). We removed nine incomplete entries, each missing a single predictor value, so our final dataset has 277 complete instances. Our final sample is small and imbalanced: out of 277 patients, only 29.24% (81 patients) had recurrence event(s) five years post-operation.[1]

Because all predictor variables were recorded as categorical variables, we created dummy variables for all but one category. We refer to this dataset as Dataset A. In addition, some attributes (namely Age, TumorSize, and #InvNodes) can be represented numerically. Although we wanted to stay true to the original data collection, we chose to create variations of these predictors to serve as numerical representations to reflect their quantitative nature. That is, for each group of numerical values, we replaced the value for all patients in that group with the mid-range (e.g. people in age group 40-49 receive age value 44.5)[2]. We refer to this dataset as Dataset B. We used both data sets to fit and compare models. In doing so, we determined the best approach for handling the qualitatively-recorded numerical data while looking for the best discriminative model.

[1] We considered filling in missing entries by predicting their values using the other predictor variables present in the dataset. However, doing so would only improve imbalance in the data marginally: the percentage of patients to have had recurrence event(s) within five years post-operation would only increase from 29.24% to 29.72%. As such, we decided to simply remove data with missing values.

[2] Note that for other categorical predictors, dummy variables were still created.

Methods and Results

We begin by using stepwise procedures to find the best first order model for each dataset considering both AIC and BIC criteria. We compare results using both a likelihood ratio test for nested models and testing each model on independent data, after converting our predicted probabilities into binary outcomes using a decision threshold of 0.5. We compute results using leave-one-out cross validation and consider model accuracy, false negative rate (FNR) and Area Under the Curve (AUC) for performance evaluation.

Because our data is medical, in our evaluation we prioritize FNR and AUC over accuracy. Our goal is not to find the model that can predict well on most instances, but rather to find one that can best distinguish patients who will have recurrence events from others. As such, we especially seek a low FNR to reduce the number of patients who are falsely predicted to *not* experience recurrence.

Table 2. Comparison between first-order models using Dataset A and Dataset B.

	(Model I)	(Model II)		(Model III)	(Model IV)
Dataset A	AIC Criterion	BIC Criterion	**Dataset B**	AIC Criterion	BIC Criterion
(Intercept)	-2.1499 ***	-1.8458 ***	*(Intercept)*	-2.91190 ***	-1.98542 ***
PostMeno	0.4698		*PostMeno*	0.50369	
Deg-Malig$_2$	0.1667	0.2940	*Deg-Malig*$_2$	0.18690	0.38510
Deg-Malig$_3$	1.5781 ***	1.6135 ***	*Deg-Malig*$_3$	1.49404 ***	1.62452 ***
Radiation	0.6185 .		*Radiation*	0.58945 .	
NodeCaps	0.8924 *	1.0807 **	*TumorSize*	0.02566 .	
			#InvNodes	0.10088 *	0.13267 **
Deviance	284.56	290.72	Deviance	281.70	334.78
(df)	(271)	(273)	(df)	(270)	(276)
AIC	**296.56**	298.72	AIC	**295.70**	299.51
BIC	318.30	**313.21**	BIC	321.07	**314.01**
Accuracy	0.7365	**0.7653**	Accuracy	0.7545	**0.7726**
FNR	**0.6667**	0.7160	FNR	**0.6296**	0.6667
AUC	**0.7026**	0.6252	AUC	**0.7344**	0.6480

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Table 2 shows the best first-order models resulting from Datasets A and B respectively. For Dataset A, after conducting a likelihood ratio test[3] and comparing FNR and AUC, we select Model I as the better model. For Dataset B, we choose Model III as the better model, based on similar likelihood ratio test results[4]. We further refine Models I and III by performing stepwise procedures with all two-term interactions between all variables found in the best first-order models. Table 3 shows the best refined models determined from initial models Models I and III. For Dataset A, both the AIC and BIC criteria give the same refined model. For Dataset B, we see that the more parsimonious Model VII outperforms Model VI in most performance metrics, including FNR. Thus, we select Model VII as the better refined model.

Although we fit Models V and VII on different datasets, neither model includes the variables that differ between the datasets (age, TumorSize, and #InvNodes), and so the fit of Models V and VII will remain unchanged across the two datasets. Thus, we can compare Model V and VII directly and conclude that Model VII is the best model for our purposes with its lower FNR and higher accuracy. We further study the residual deviances of Model VII and find no influential outliers[5] We also check for severe multicollinearity issues with a VIF threshold of 5 and find that there isn't a problem.

[3] Under the null hypothesis $H_0 : \beta_{PostMeno} = \beta_{Radiation} = 0$ with $\alpha = 0.05$, our observed test statistic is 6.16 $\sim \chi^2(2)$. Thus, we reject the null hypothesis and conclude that, with a p-value of 0.0459, at least one of these predictors is significant.

[4] Under the null hypothesis $H_0 : \beta_{PostMeno} = \beta_{Radiation} = \beta_{TumorSize} = 0$ with $\alpha = 0.05$, our observed test statistic is 9.81 $\sim \chi^2(6)$. Thus, we reject the null hypothesis and conclude that, with a p-value of 0.2026, at least one of these predictors is significant.

[5] To do so, we remove each observation from the model and investigate the change in magnitude in residual deviance as a result of the removal.

Table 3. Comparison between refined models using Dataset A and Dataset B.

	(Model V)		(Model VI)	(Model VII)
Dataset A	AIC & BIC	**Dataset B**	AIC Criterion	BIC Criterion
(Intercept)	-2.5136 ***	*(Intercept)*	-5.18617 ***	-2.5402 ***
PostMeno	1.0796 **	*PostMeno*	1.10716 **	1.0896 **
Deg-Malig$_2$	0.1057	*Deg-Malig = 2*	2.55637 .	0.2872
Deg-Malig$_3$	1.6699 ***	*Deg-Malig = 3*	3.47500 *	1.9403 ***
Radiation	1.9090 ***	*Radiation*	1.84538 **	2.1151 ***
Node_Caps	0.9143 *	*TumorSize*	0.09145 *	
		#InvNodes	0.10108 *	
*PostMeno*Radiation*	-2.1454 **	*PostMeno*Radiation*	-2.00544 **	-2.1167 **
		Deg-Malig$_2$**TumorSize*	-0.08467 .	
		Deg-Malig$_3$**TumorSize*	-0.06504	
Deviance	274.86	Deviance	268.55	281.04
(df)	(270)	(df)	(267)	(271)
AIC	288.86	AIC	**288.55**	293.04
BIC	314.23	BIC	324.79	**314.79**
Accuracy	0.7473	Accuracy	0.7509	**0.7726**
FNR	0.6049	FNR	0.6049	**0.5926**
AUC	0.7019	AUC	**0.7319**	0.6824
Signif. codes:	0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1			

Conclusions and Other Considerations

Here we built a predictive and discriminative tool for breast cancer recurrence by building a logistic regression model. Our final fitted model is: $\widehat{\text{logit(recurrence)}} = -2.54 + 1.09 \cdot \text{PostMeno} + 0.29 \cdot \text{Deg\_Malig}_2 + 1.94 \cdot \text{Deg\_Malig}_3 + 2.12 \cdot \text{Radiation} - 2.12 \cdot \text{PostMeno*Radiation}$. To a large extent, the included predictors are as expected. As for menopausal status and radiation, a post-menopausal woman's predicted odds of recurrence decrease if they had radiation. This is likely because a woman who has already undergone menopause may have caught the cancer earlier as post-menopausal women are at higher overall risk due to greater cumulative estrogen exposure [8]. Further, degree of malignancy is associated with cancer stage--a lower degree of malignancy, a lower cancer stage and thus a lower odds of recurrence. It is also likely that radiation treatment positively correlates to disease severity since this treatment is to destroy cancer cells. Specifically, doctors justify this treatment only when the tumor is not already widespread (i.e. in late stage cancer cases).

Still, our model and its implications are limited by nature of the data and our knowledge of its background. Since 1988, the year the data were collected, breast cancer treatments have changed in ways that specifically help particular individuals lower their cancer risk and recurrence rate. For example, in 2006 the drug raloxifene was approved to reduce risk in postmenopausal women [9]. Further, though we know recurrence status five years following initial diagnosis and that in the case of breast cancer this is within the window of time during which most recurrence occurs, we are not aware if recurrence status changes later. It might be interesting to look into whether younger individuals experience recurrence at a later age, though age is insignificant in our model. Because we don't know how the data were collected, moreover, we lacked exact values when building our model with numerical predictors and our mid-range value may have been an inaccurate estimation for certain individuals.

With these limitations in mind, we hope to extend this project in a few directions. First, we want to experiment with various decision thresholds in efforts to reduce the false negative rate--perhaps at a small cost to overall model accuracy. We also wish to explore and potentially improve our model by supplementing it with more balanced and modern data. By doing so, we believe our model can likely be extended to predict the number of recurrence events or even to predict various survival rates given similar predictors to those found in our model.

References

[1] Centers for Disease Control and Prevention. (2017). Basic Information About Breast Cancer. Retrieved from https://www.cdc.gov/cancer/breast/basic_info/index.htm

[2] Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M. C., Dossus, L., & Rondeau, V. (2018). Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. BMC cancer, 18(1), 171.

[3] Susan G. Komen For the Cure. (2018). Breast Cancer Recurrence.
Retrieved from https://ww5.komen.org/BreastCancer/ReturnofCancerafterTreatment.html

[4] World Cancer Research Fund International. (2015). Breast Cancer Statistics. Retrieved from https://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics

[5] American Cancer Society. (2016). Treatment of Recurrent Breast Cancer. Retrieved from https://www.cancer.org/cancer/breast-cancer/treatment/treatment-of-breast-cancer-by-stage/treatment-of-recurrent-breast-cancer.html

[6] Dua D. & Karra Taniskidou, E. (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

[7] Zwitter, M. & Soklic, M. Breast Cancer Data. (1988). Institute of Oncology, University Medical Centre Ljubljana, Yugoslavia.

[8] Nathan-Garner, L. (2015, November). How does menopause affect cancer risk? MD Anderson Cancer Center. Retrieved May 14, 2018, from https://www.mdanderson.org/publications/focused-on-health/november-2015/FOH-menopause-cancer.html

[9] Barrett-Connor, E., Mosca, L., Collins, P., Geiger, M. J., Grady, D., Kornitzer, M., ... & Wenger, N. K. (2006). Effects of raloxifene on cardiovascular events and breast cancer in postmenopausal women. New England Journal of Medicine, 355(2), 125-137.