Classification Model Using Basic Patient Information and Bloodwork to Predict Breast Cancer

Tomas Meade

Abstract

Introduction

This paper focuses on creating and assessing a logistic regression classification model to predict the risk of breast cancer using easily attainable patient information and bloodwork samples.

Methods

The variables in the dataset consist of basic patient information and blood work analysis for all 116 observations. The variables are the age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin (ng/mL), MCP-1(pg/dL) and a variable that classifies if a participant has breast cancer or not. Logistic regression was used to create the model and it was assessed by calculating the concordance index, sensitivity and specificity.

Results

The final logistic regression model consisting of BMI, glucose, insulin, HOMA, resistin and MCP-1 had concordance index of 0.73. At a threshold value of 0.50, the sensitivity was 59% and the specificity was 74%. Lowering the threshold value to 0.44 achieved a sensitivity of 74% and a specificity of 53%.

Conclusion

These results suggest that there is some evidence to suggest that models using easily attainable parameters could have potential in predicting the risk of breast cancer in women.

Introduction

This project sought to evaluate if easily attainable patient information and data which can be gathered in routine blood analysis are effective in predicting breast cancer. The outcome of interest was the binary variable categorizing whether a patient has breast cancer or not. The predictor variables are age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP-1. The main reason these variables were chosen is because they are attained easily and inexpensively while also having ties to breast cancer. The age and BMI of a patient are very basic patient data that a hospital would have on record for most likely all their patients. The rest of the variables can be obtained in a routine blood analysis. These variables have also been linked to breast in previous literature. So, overall the variables were chosen by balancing their attainability and their known association with breast cancer.

Methods

The dataset consisted of 116 observations made up of 64 cancer patients and 52 healthy controls. The source of the data was a study done at the Gynaecology Department of the University Hospital Centre of Coimbra (CHUC) between 2009 and 2013. Women who were recently diagnosed with breast cancer at CHUC were recruited and all samples were collected before surgery and treatment. Healthy females were then also recruited and enrolled in the study as controls. All participants were free from any infection or other acute diseases or comorbidities at the time of enrollment in the study and never received any prior cancer treatment. The variables in the dataset consisted of basic patient information and blood work analysis for all 116 observations. The basic patient information was the age (years) and BMI (kg/m2) of each participant and a variable that classifies whether a participant has breast cancer or not. The bloodwork information consists of glucose (mg/dL), insulin (µU/mL) HOMA ((glucose in mmol/L x insulin in mIU/mL)/22.5), leptin (ng/mL), adiponectin (µg/mL), resistin (ng/mL), and MCP-1(pg/dL). All blood samples were collected at the same time of the day after an overnight fasting.

To begin the analysis, box plots were created in order to examine the relationship between some of the predictor variables and the outcome variable which classifies whether a patient has breast cancer or not. Then the data was split up randomly into training and test groups with 60% of the data in the training set and 40% of the data in the test set. The variables and their relationship to the outcome variable were examined individually in single variable logistic regression models. Variable selection was then done with the training dataset using the best subsets method to determine the optimal set of predictor variables. The criterion used to determine the optimal subset was the BIC value of each model. With the final logistic model constructed, the c-index was calculated on the test data set to measure the concordance rate for the model. Then both the sensitivity and specificity of the model were calculated using the test set with various values of the threshold value c. The value of c was analyzed by focusing on a higher sensitivity rate in order to prioritize identifying anyone at risk of having breast cancer. This was done by creating an ROC curve to visualize the trade off between the sensitivity and specificity rate.

Results

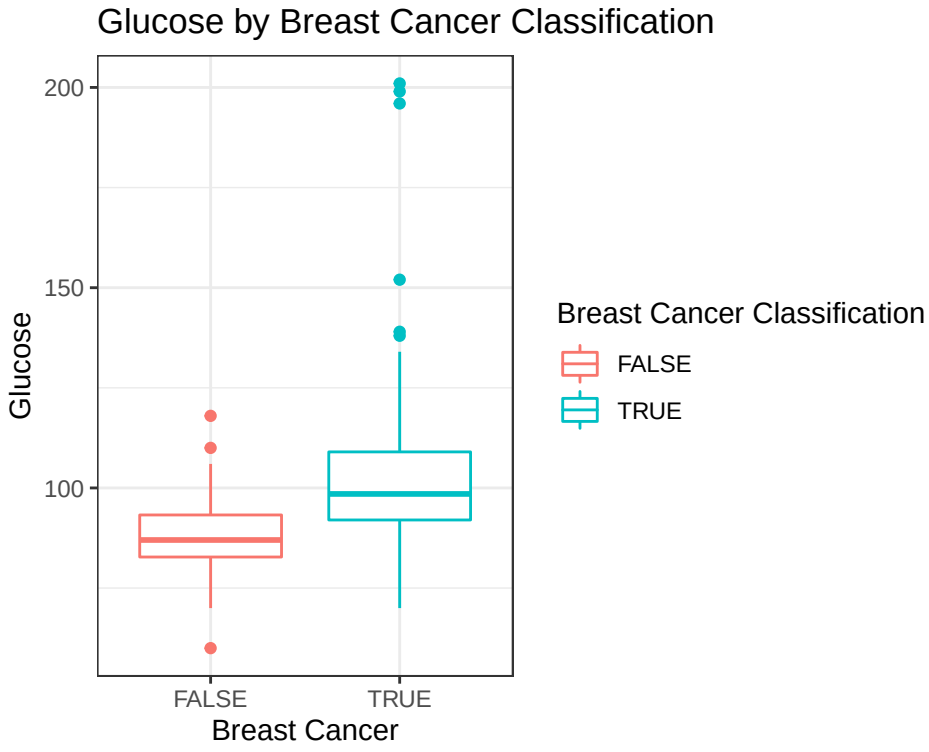

Figure 1: Boxplot showing the observed difference between glucose levels of healthy controls versus breast cancer patients.

From the basic exploratory analysis, the glucose, HOMA and insulin levels of healthy controls versus breast cancer patients seemed to have the most obvious observed difference. This was confirmed in the single variable logistic regression models since glucose, HOMA and insulin had the strongest association to breast cancer with p values that would be significant at any reasonable alpha-level (see table 1).

The final logistic regression model consisted of five parameters. The final set of variables were BMI, glucose, insulin, HOMA and resistin. Using an alpha-level of 0.05, each variable was significant (see table 2). The concordance index of the final model was 0.73. At a threshold value of 0.50, the sensitivity was 59% and the specificity was 74%. To prioritize sensitivity, lowering the threshold value to 0.44 achieved a sensitivity of 74% and a specificity of 53%. The trade off between sensitivity and specificity is visualized in the ROC curve in figure 2.

The concordance index of 0.73 shows some indication that the model was effective in discriminating between healthy controls and breast cancer patients. With a threshold value of 0.44, the model shows potential to

have a strong true positive rate correctly identifying participants with breast cancer while still maintaining a true negative rate of over 50%.

Table 1: Each row represents a logistic regression model with the stated predictor variable and the outcome variable which classifies whether a patient has cancer or not. Each predictor variable is listed along with its associated coefficient and p-value.

Variable	Coefficient	p-value
Age	-0.00135	0.642
Adiponectin	-0.00142	0.835
BMI	-0.01319	0.156
Glucose	0.00852	2.05e-05
HOMA	0.03895	0.002
Insulin	0.01373	0.003
Leptin	-0.00003	0.991
Resistin	0.00916	0.014
MCP1	0.00013	0.329

Table 2: Each predictor variable is listed along with its associated coefficient and p-value in the final logistic regression model.

Variable	Coefficient	p-value
BMI	-0.02587	0.01256
Glucose	0.01348	4.36e-05
Insulin	0.05390	0.00032
HOMA	-0.15708	0.00109
Resistin	0.01014	0.04641

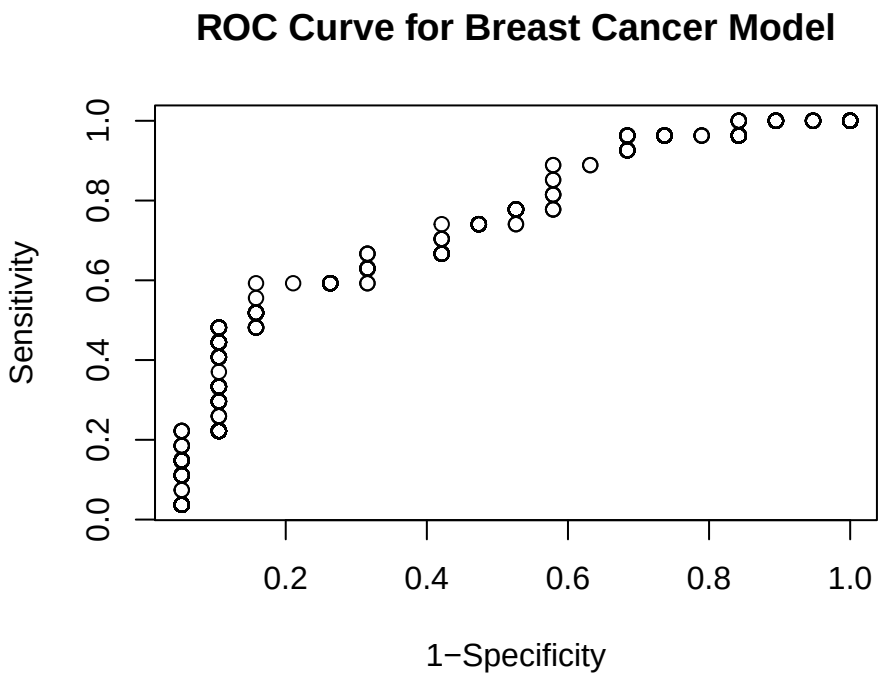

Figure 2: ROC curve showing the trade of between sensitivity and specificity.

Discussion

These results are mostly consistent with previous work, but other studies have resulted in more promising results. The original study done at CHUC constructed models that achieved sensitivity and specificity rates over 80%. This is most likely because the study done at CHUC was more in depth and utilized a better understanding of the variables involved in the model. This paper uses logistic regression and the best subsets method to construct the final model without an extensive knowledge of the biology behind the parameters and their association with breast cancer. The BIC value was used as a criterion for variable selection in order to emphasize the theme of developing a simple model with only a small amount of parameters and to penalize overfitting. However, using the BIC value could have penalized having more parameters to a degree that was too extreme, resulting in a over simplified model. The results of this paper are still promising since they reaffirm the potential for the effectiveness of using simple parameters to help predict the risk of breast, but they are not entirely conclusive.

Conclusion

This paper conducts a brief statistical analysis of the predictive capability of a logistic regression model to determine the risk of breast cancer. The analysis was done without an extensive knowledge of the parameters of interest or breast cancer. However, the results do align with previous literature in the field and suggest that only anthropocentric and bloodwork data can still be effective in predicting the risk of breast cancer. Further studies should be done to assess these results and whether or not they could be applied to the general population.

References

Assiri AM, Kamel HF. Evaluation of diagnostic and predictive value of serum adipokines: Leptin, resistin and visfatin in postmenopausal breast cancer. Obes Res Clin Pract. 2015;10(4):442-53.

Cole KD, He HJ, Wang L. Breast cancer biomarker measurements and standards. Proteomics Clin Appl. 2013;7(1–2):17–29.

Crisóstomo J, et al. Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer. Endocrine. 2016;53(2):433-42.

Kloten V, et al. Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based breast cancer screening. Breast Cancer Res. 2013;15(1):R4.

Patrício, M., Pereira, J., Crisóstomo, J. et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 18, 29 (2018).