

An Application of Generalized Linear Models to Fine Needle Aspiration in Breast Cancer

Chudi Zeng*

SDU-ANU Joint Science College, Shandong University, Weihai 264209, China

*Corresponding Author Email: 202000700258@mail.sdu.edu.cn

Abstract. Breast cancer is currently the most dangerous cancer for women worldwide. Doctors routinely employ biopsies, diagnostic mammograms, and other techniques to detect and diagnose breast cancer. The Fine Needle Aspiration, also referred to as fine needle biopsy, is a technique for assessing tumors that involves inserting a needle into a mass to retrieve alive cells. However, the current breast biopsy test is time-consuming and unable to detect early breast cancer. Applying the statistical tools to fine-needle aspiration is helpful in developing its feasibility and reducing test time, thereby reducing the cost of service as well as waiting time. In this study, the diagnostic model was fitted with a generalized linear model as the framework and Least Absolute Shrinkage and Selection Operator regression as the essential methods. Amongst cellular level features, which are variables in the model, some features were identified that play an essential role in the models, including texture, smoothness, concave points, and fractal dimension. The high accuracy (>0.9) obtained from the model in data testing supported that Generalized-Linear-Models-based machine prediction can effectively assist physicians in their clinical diagnosis. In addition, essential features in the model could be considered to have some association with the hidden lesion of breast cancer.

Keywords: breast cancer; fine-needle aspiration; generalized linear models; LASSO; prediction.

1. Introduction

Nowadays, breast cancer has become the most lethal kind of female cancer, which ranks the first in incidence in 85% of countries and first in mortality in 60% of countries [1]. Ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS), and triple-negative tumors are recognized as the three main categories of breast cancer. Breast cancer is classified according to two molecular targets: estrogen receptor alpha ($ER\alpha$) and epidermal growth factor 2 (ERBB2, formerly HER2 or HER2/neu). The incidence, prognosis, and systemic treatment options vary among the three subtypes of breast cancer [2]. Family history and genetic susceptibility, breast density in the clinical mammographic images, and history of atypical ductal hyperplasia or atypical lobular hyperplasia are all risk factors for DCIS as well as LCIS [3].

A popular and effective option for breast cancer screening is mammography, including magnetic resonance imaging, tomography, and ultrasound. However, this is not recommended because it may increase the risk of cancer for women [4]. New techniques based on improved mammography will help balance the effectiveness of mammography with the hazard, but there is no denying that the danger cannot be eliminated entirely.

Fine needle aspiration cytology (FNAC) is a minimally invasive approach to tumor sampling with almost few risks [5]. FNAC is reliable in detecting HPV16 DNA in neck masses and provides strong evidence for the diagnosis of HPV-positive oropharyngeal squamous carcinoma. It also offers morphological and immunophenotypic features to confirm the diagnosis of pancreatic cancer [6]. Nina et al. described the characteristics of some common and rare FNAC samples in the breast to help for the application of FNAC in clinical diagnosis [7]. In contrast to mammography, which can only be diagnosed when lesions are evident, FNAC may allow earlier detection of potential lesions.

As a classic statistical tool, generalized linear models (GLM) can be used in the life sciences to reflect an organism's behavioral or physiological characteristics [8]. The GLM maps the relationship between lesion occurrence and risk variables (e.g., hypertension) in white matter lesions, showing the accuracy and computational efficiency of the method [9]. Given that, it is supposed that the GLM can

do the same when analyzing the FNA of breast cancer to help the physician identify the lesion and make the diagnosis.

This study aims to examine the possibility of applying statistical models in biopsy procedures and clinical studies. We modeled the available breast cancer FNA data by GLM. We describe the distribution of different cellular features in the model and finally briefly evaluate the research value of significant features according to the model accuracies.

2. Methods

2.1. Data collection

The data set was sourced from the UCL Machine Learning Repository. This study used 569 observations from the Breast Cancer Wisconsin Diagnostic Data Set and 198 observations from the Breast Cancer Wisconsin Prognostic Data Set [10].

The features collected were radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension, diagnosis, outcome, and time. Given that the characteristics were assessed using digitized imaging of fine-needle aspirates from breast lumps, they described the properties of the cell nuclei depicted in the images. The independent variables are the mean, standard error, and maximum discrepancy values of all nuclei features. Patients are classified according to whether the cancer is benign or malignant or whether it has recurred.

2.2. Model selection

2.2.1 The Akaike information criterion (AIC)

Akaike's Information Criteria (AIC) is a robust and commonly used model selection criterion [11]. The AIC formula is as follows:

$$AIC = 2K - 2 \ln(L)$$

When we have the AIC scores of several models, we select the model with the lowest AIC score as the optimum model. A lower AIC value shows that the model requires less information to predict with almost the same degree of accuracy [12]. However, the best-approximated model given by the AIC is not always proportional in biological significance to statistical significance. The AIC was used for the initial filtering, and the final selected model will be derived by combining biological significance and other judgment criteria.

2.2.2 Variance Inflation Factor

The Variance Inflation Factor (VIF) is a commonly used method for analyzing the presence of a high degree of collinearity in multiple linear regression models. The VIF formula is as follows:

$$VIF = \frac{1}{1 - R^2}$$

R^2 is the degree of correlation between the predictor and the other predictor variables in the linear regression. A VIF greater than ten will usually be considered a high degree of collinearity, WHICH means that the conclusion of the analysis may be questioned [13]. For variables with too high VIF, we choose to solve this problem by selectively removing them because of the redundancy of the information provided by these variables.

2.3. Model building

2.3.1 GLM

A series of p predictor variables and a response variable are used in ordinary multiple linear regression to fit a model of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

Y_i : The response variable

X_i : The j^{th} predictor variable

β_j : The average effect on Y_i of a one-unit increase in X_{ji} , holding all other predictors fixed

ϵ_i : The error term

Fitted models are framed by a family of GLM and may be used to examine the strength of the correlation between particular variables and breast cancer diagnosis or recurrence.

2.3.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression is an appealing alternative since it tackles both overfitting and overestimation of the model's performance in terms of utilizing the included variables to explain observed variability ('optimism bias') [14]. The mathematical expression for the LASSO criterion is:

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \dots - \beta_p X_{pi})^2 + \lambda \sum_{j=1}^p |\beta_j|$$

λ is adjustment parameters, the larger λ penalizes the linear model with more variables more strongly. $\frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \dots - \beta_p X_{pi})^2$ is the Mean Square Error (MSE) and $\lambda \sum_{j=1}^p |\beta_j|$ is the penalty for model complexity. LASSO, which can constrain the model's parameters, calculates the model coefficients with a limited prediction error by cdenyingthe sum of the regression coefficients' absolute values to be less than a λ .. In contrast, even some regression coefficients will shrink to 0.

2.3.3 Cross-validation (CV)

CV is available to encourage selecting a model that yields a trustworthy calibration from the general sample [11]. In real-world applications of regression, one may naturally consider using a CV to choose a parametric estimator to pursue a better estimation accuracy. CV will partition the observations randomly into ten non-overlapping groups/folds of equal size. For instance, the validation set will be the first fold, and the model will fit nine times. The bias-variance advantages are frequently given as a major for using such model validation methodologies. For LASSO, CV helps to choose the tuning λ parameter [15].

3. Results

3.1. Wisconsin Diagnosis of Breast Cancer

To fit a GLM for prediction, a mixture of stepAIC, LASSO regression, and CV was applied. VIF and AIC can assist in judging the merits of the model. While high accuracy levels are achieved, models with fewer variables are preferentially adopted.

Table 1. Model information from generalized linear regression at Wisconsin Diagnosis of Breast Cancer

Model	Number of variables	Prediction accuracy	AIC	Residual Deviance
All variables	10	0.9848649	99.97	79.97
Mean variables	3	0.9400541	172.3	164.3

First, all 30 variables were called to fit the logit linear model, and then stepAIC was performed to improve the model's accuracy. CV and LASSO were integrated for the next step of model fitting. By omitting individual variables with the criterion of VIF more outstanding than 10, the first model was obtained. Considering the implications of different variables, which are various cell nuclear features, on breast cancer diagnosis from Wisconsin, an additional repeat the above steps was performed by selecting only all mean variables. Here the information from the conditional fitting of the model is

shown in Table 1. The accuracy values are 0.9848649 and 0.9400541, indicating solid predictive models.

3.2. Wisconsin Prognostic of Breast Cancer

Table 2. Model information from generalized linear regression at Wisconsin Prognostic of Breast Cancer

Model	Number of variables	Prediction accuracy	AIC	Residual Deviance
All variables	8	0.7739	220.9	202.9
Mean variables	5	0.7736	215	203

The processing of the Prognostic dataset is approximately the same as that of the Diagnosis dataset. Given the differences in data and the functional purpose of the model, the penalization of LASSO needs to be adjusted to avoid over-fitting. Here the information from the conditional fitting of the model is shown in Table 2. The accuracies of the models fitted to the prognostic data set are 0.7739 and 0.7736. The exactness of the prognostic model was lower than that of the diagnostic model, suggesting that the prognostic model should play a significantly more minor role in the subsequent analysis than the diagnostic model.

3.3. Model comparison

Table 3. Comparison of variable parameters from different models

	Diagnosis		Prognostic	
Model	All variables	Mean variables	All variables	Mean variables
Intercept	-45.915	-21.138	3.223	0.339
Texture (m)	0.239	0.326	-0.016	-0.055
Smoothness (m)	73.989	/	40.237	37.243
Concavity (m)	/	/	4.247	/
Fractal dimension(m)	/	/	-90.928	-45.888
Texture (s)	/	/	-0.672	/
concave points (s)	/	/	-77.987	/
Symmetry (s)	/	/	39.348	/
Symmetry (w)	11.312	/	-6.243	/
Area (m)	/	/	/	0.001
Symmetry (m)	/	/	/	-12.633
concave points (m)	52.0754	100.892	/	/
Smoothness (s)	274.655	/	/	/
Fractal dimension (s)	-844.599	/	/	/
Radius (w)	1.365	/	/	/
Texture (w)	0.161	/	/	/
Concavity (w)	8.506	/	/	/
Radius (m)	/	0.655	/	/

Notes: “m,” “s,” and “w” mean the average, standard error, and the maximum deviation (the average of three times) of the data for that observation for a large number of features of this cell nucleus, “/” means that this variable is not present in the model.

A comparative study of the results from these two datasets may uncover the causes of breast cancer development at the level of cell nuclear features. The information on the model variables with different data sets is shown in Table 3. Although the texture is not a significant parameter, it appears in all models to show that it is an integral part of the models. According to smoothness appears in both diagnostic and prognostic models, and whose parameters are much larger than texture parameters, it is supposed that smoothness is also associated with breast cancer lesions. Meanwhile,

although the concave points only appeared in the diagnostic model, especially the mean model, it has a significant coefficient, so they can also be considered as a relative factor with a breast cancer lesion. At the same time, the fractal dimension is the opposite of the concave point and appears only in the prognostic model.

4. Discussions

Both datasets were regressed by stepAIC and LASSO with cross-validation to yield a model containing all kinds of variables and a model containing only “mean” variables, whereby we delivered four models in total. Among them, the models about diagnosis had prediction accuracies of 0.9848649 and 0.9400541. We can use the model to perform computations on fine-needle aspirates of breast masses and determine whether the tested person's breast cancer is benign or malignant. However, the accuracy of 0.7739 and 0.7736 does not seem to allow the other two models to be used directly in clinical practice to determine whether the tested person will have breast cancer recurrence. Although the models on Prognostic cannot be applied in practice, they are still instructive for theoretical studies. It would be a great achievement if the data obtained from the analysis of fine-needle aspirates alone could determine the recurrence rate and five-year survival rate.

Due to the high accuracy of the model predictions and the fact that the model's variables are the nucleus characteristics of breast cells, it can be considered that the variables that appear in all the crowded models may have some relevance to the etiology of breast cancer. In statistics and biology, the “mean” variable is able to characterize this cellular nuclear feature of the observation. Texture, smoothness, concave points, and fractal dimension features appear most frequently (Table 3) and are essential in the crowd variables. The concave points did not appear in the prognostic model. Still, this prognostic model's accuracy is not convincing enough to confirm that there is no association between the notch and the recurrence of breast cancer. A similar conclusion can be drawn for the fractal dimension, which is the opposite of the concave points' case. The associations between concave points and reproduction results, fractal dimension, and diagnostic results need further exploration. There is mounting evidence that textural features may be helpful to in predicting breast cancer risk, which is integrated into breast cancer risk assessment models [16]. Texture and smoothness are also used in the Synchronous Breast Cancer study on Phenotypic Similarities on MRI [17]. Thus, this is an effective approach to finding lesions, as proven by the research in previous papers.

The model's strength is that, as the generalized linear regression model, including all kinds of Prognostic models, all models had a significance F value in Variance Analysis (ANOVA) less than 0.05. This derives the P-value (less than 0.05) from testing the null hypothesis that the data for all groups are from groups with the same mean, indicating that this is a robust predictive model as measured by statistical analysis of variance.

There are still some limitations in this study. The models' data are too old and may be unsuitable for today's situation, especially because coronavirus disease will affect people's immunity and physical fitness. In clinical application, newly updated data is needed to create a model that fits the current situation well. In that case, more new observations are required to make adjustments to the parameters in the model. Similarly, the training method of the model on prognostic needs to be changed to improve its prediction accuracy in future studies. In addition, exact information about the process of sample collection is unknown, which may also influence the prediction results, as keeping the precise specifications in operation at the sampling time would have resulted in more accurate data.

5. Conclusion

Analysis of breast cancer data with GLM will yield good statistical significance, especially in pathological exploration and diagnostic prediction. This study fitted a series of models with methods such as LASSO regression and cross-validation, which applied the generalized linear model as the

framework. The model got high predictive accuracy (≥ 0.94), thereby having potential for clinical application.

A comparison of the models identified four variables with the most commonality and importance, including texture, smoothness, concave points, and fractal dimension. Since the model variables were statistically significant ($p\text{-value} < 0.05$), these features can be considered to be statistically associated with tumor formation or progression, which is consistent with previous studies. Hence, it can be supposed that the combination of FNA and GLM is helpful for researchers in lesion exploration. In the future, the introduction of statistical models for a comprehensive interpretation of big data may provide mathematical help for cancer research. By improving algorithms or replacing new data, the development of increasingly accurate and testable models for disease research objectives is within reach.

References

- [1] H. Sung et al. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.*, 71: pp. 209–249.
- [2] A. G. Waks, E. P. Winer. (2019) Breast Cancer Treatment: A Review. *JAMA*, 321: p. 288.
- [3] E. M. Ward et al. (2015) Cancer statistics: Breast cancer in situ: Cancer Statistics: Breast Cancer in Situ. *CA. Cancer J. Clin.*, 65: pp. 481–495.
- [4] C. E. Comstock et al. (2020) Comparison of Abbreviated Breast MRI vs Digital Breast Tomosynthesis for Breast Cancer Detection Among Women With Dense Breasts Undergoing Screening. *JAMA*, 323: p. 746.
- [5] T. Ahmad, M. Naeem, S. Ahmad, A. Samad, A. Nasir. (2008) Fine needle aspiration cytology (FNAC) and neck swellings in the surgical outpatient. *J. Ayub Med. Coll. Abbottabad JAMC*, 20: pp. 30–32.
- [6] L. Sivars et al. (2017) Human papillomavirus DNA detection in fine-needle aspirates as indicator of human papillomavirus-positive oropharyngeal squamous cell carcinoma: A prospective study. *Head Neck*, 39: pp. 419–426.
- [7] N. S. Shabb, F. I. Boulos, F. W. Abdul-Karim. (2013) Indeterminate and Erroneous Fine-Needle Aspirates of Breast with Focus on the “True Gray Zone”: A Review. *Acta Cytol.*, p. 16.
- [8] A. Calabrese, J. W. Schumacher, D. M. Schneider, L. Paninski, S. M. N. Woolley. (2011) A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PloS One*, 6: p. e16104.
- [9] P. Kindalova, I. Kosmidis, T. E. Nichols. (2021) Voxel-wise and spatial modelling of binary lesion masks: Comparison of methods with a realistic simulation framework. *NeuroImage*, 236: p. 118090.
- [10] Dua, D., Graff, C. (2019). UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>.
- [11] J. Ranstam, J. A. Cook. (2018) LASSO regression. *Br. J. Surg.*, 105: Art. no. 10.
- [12] W. Pan. (2022) Akaike’s Information Criterion in Generalized Estimating Equations. *Biometrics*, 57: p. 7.
- [13] R. Salmerón, C. B. García, J. García. (2018) Variance Inflation Factor and Condition Number in multiple linear regression. *J STAT COMPUT SIM*, 88: p. 21.
- [14] D. C. Camilo, L. Lombardo, P. M. Mai, J. Dou, R. Huser. (2017) Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized Generalized Linear Model. *Environ. Model. Softw.*, 97: pp. 145–156.
- [15] Y. Yang. (2007) Consistency of cross validation for comparing regression procedures. *Ann. Stat.*, 35: Art. no. 6.

- [16] A. Gastounioti, E. F. Conant, D. Kontos. (2016) Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. BCR, 18: p. 91.
- [17] H. Wang, B. H. M. Velden, H. S. M. Chan, C. E. Loo, M. A. Viergever, K. G. A. Gilhuijs. (2020) Synchronous Breast Cancer: Phenotypic Similarities on MRI. J. Magn. Reson. Imaging, 51: pp. 1858–1867.