# Logistic Regression Analysis of Breast Cancer From Mammographic Evaluation

Parviz Abdolmaleki[1], Ph.D., Masomeh Gity,M.D[2], Masomeh Tahmasebi[1], M.Sc, Majid Rohandeh[1], M.Sc.

**1.Department of Biophysics , Tarbiat Modares University, Tehran.**

e-mail: parviz@modares.ac.ir

Abstract: Logistic regression analysis is used to differentiate malignant from benign in a group of patients with proved breast lesions on the base of morphological data extracted from the conventional mammogram. Our database include 122 patients' records consisting 12 qualitative variables. The database is randomly divided into the training and validation samples including 82 and 40 patients' records respectively. The training and validation samples are used to construct the logistic regression modelas a classifier and to validate its performance respectively. Finally, important criteria such as sensitivity, specificity, accuracy and receiver operating characteristic curve (ROC) analysis for this method as well as that of the radiologist are compared. Our results show that the logistic regression model is able to classify correctly 31 out of 40 cases presented in the validation sample. Comparing the output of this method with that of the radiologist shows a reasonable diagnostic accuracy 78%, a high specificity (81%) and a moderate sensitivity (72%).

Keywords: Breast cancer, Logistic regression analysis, ROC curve.

## 1- Introduction:

Breast cancer is the first cause of cancer deaths among women. Mammographic screening can reduce the mortality from breast cancer by as much as 20–30% [1,2], because it allows detection of non-palpable, non-invasive and early invasive tumors. Approximately 35% or less of women who undergo biopsy for histopathologic diagnosis of breast cancer are found to have malignancies [1,2]. One goal of the application of computer-aided diagnosis (CAD) to mammography is to reduce the false-positive rate. Avoiding benign biopsies spares women unnecessary discomfort, anxiety, and expense.

Since the final histologic diagnosis being benign (with probability of p) or malignant (with probability of (1-p)) as a binary outcome, the logistic regression model [3] could be used as a CAD system in form of a classifier to predict the outcome of biopsy. This is a form of regression model which is used when the dependent variable is a dichotomy and the independent variables are continuous, categorical or both. Logistic regression model has been successfully performed for different computational problems in pattern recognition and decision making [4,5].

In this study we intend to establish a logistic regression model to work as a tool for radiologist to predict the outcome of biopsy using data extracted from the conventional mammography. The performance of the established model is then compared with that of radiologist using the common statistical indices including accuracy, sensitivity, specificity and receiver operating characteristic curve (ROC) analysis.

2- **Materials and Methods:** Our goal was to apply the logistic regression analysis to

the data collected in a study designed to predict the malignancy of breast cancer on the basis of features that had been extracted from the conventional mammogram using defined criteria. Our study group consists of 122 consecutive patients (age ranged 23-80 years; mean age, 49.7 years) with histopathlogically proof. The patient group included 51 malignant lesions and 71 benign entities.

2-1. Data acquisition: The imaging was performed at the center of imaging of the Imam Khomainei Hospital during 2000 to 2002. Hook wire localization of the microcalcifications under mammographic guidance was used in all cases. Two patients had three clusters and nine patients had two clusters of microcalcifications on the mammogram of the same breast and seven patients had one cluster in each breast. The remainder patients had only one cluster of microcalcifications in either breast. All lesions were histologically confirmed after biopsy or surgical excision. An expert radiologist read the mammogram images (figure 1) and graded his finding on the following features: mass size, shape, margin, density, asymmetric density, parenchymal distortion, calcification size, shape, number, density, distribution, general impression of the radiologist based on the microcalcification data and associated features. The presence of associated features was ranked on a scale of 0-4 with increasing likelihood of malignancy. In the case of more than one associated feature, the one with the highest rank was considered. The findings were ranked using a 2-5 scale categorization with increasing likelihood of malignancy. Table 1 shows all the parameters in our database, which represented the subjective features extracted by participated radiologist (MG).

2.2 - Logistic regression model

Logistic regression is a statistical model for analysis of the relationship between an observed proportion (binary outcome) y and a vector $\mathbf{X}' = [X_1, X_2, ..., X_p]$ of regressor variables which are continuous, categorical or both for each of N individuals.

The major purpose of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predictng the response variable. Variables can be entered into the model in the order specified by the researches or logistic can test fit of the model after each coefficient is added or deleted, called stepwise regression.Cox [6] and Day and Kerridge [7] both suggested the logistic regression model for posterior probabilities as a basis for discrimination two populations $\Pi_1$ and $\Pi_2$ with prior probabilities $p_1$ and $p_2$ respectively. The objects are ordinarily separated or classified on the basis of measurements on p associated random variables $\mathbf{X}' = [X_1, X_2, ..., X_p]$. The simplest optimizing method of discrimination is to maximize the probability of correct allocation. This is achieved by allocating the sample point $\mathbf{X}$ to $\Pi_1$ (i.e. the response variable y=1) if

$$\Pr(y=1|\mathbf{X}) = \Pr(\Pi_1|\mathbf{X}) \geq \Pr(\Pi_2|\mathbf{X}) = \Pr(y=0|\mathbf{X})$$

otherwise to $\Pi_2$. Where, $p = \Pr(y=1|\mathbf{X}) = \Pr(\Pi_1|\mathbf{X})$ is given at (1) and $\Pr(\Pi_1|\mathbf{X}) + \Pr(\Pi_2|\mathbf{X}) = 1$.

The allocation of new individuals can be performed on the basis of scores given by the logit function i.e.

$$Logit(p) = ln(p/1-p) = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ,..., + \beta_p X_p)$$

If it is positive (with assumption of equal prior probabilities) the individual is allocated to $\Pi_1$ otherwise to $\Pi_2$. The logit coefficients $\beta$ are estimated by the maximum likelihood estimation (MLE) using the iterative equations (4). To test the null hypothesis that a particular logit coefficient is zero the Wald's statistics is used. This is the square of the ratio of the estimated logit coefficient to its standard error and has a chi-square distribution [4].

To establish the logistic discriminant models which could improve radiologist's performance in differentiating malignant from benign lesion, a group of patients with proved breast lesions on a base of morphological data, extracted from mammography images, is considered. Then the above-mentioned features are obtained. Using a data base consisted of 122 cases, we randomly selected two thirds (82) patients (including 49 benign and 33 malignant cases whose group identity is known) to compose the training sample. To prepare the validation sample the rest of data (40) patients (consists of 22 benign and 18 malignant cases) were selected. The logit coefficients are estimated using the Proc Logistic in SAS statistical package based on MLE method. Using the Wald's statistic at 0.05 levels, the final model for classifying new cases contains the significant independent variables .

2.3- Performance Evaluation: We applied the ROCFIT software to create the ROC curve which is a plot of sensitivity versus (1-specificity). After the logistic regression classifier had been established perfectly the validation samples was presented to the model giving two posterior probabilities. Taking into consideration the posterior probability of malignancy ($\Pr(\Pi_2 | \mathbf{x})$), the diagnostic performance of this approach was estimated. In this regard, the true positive and the false-positive fractions were determined. These data were then used to plot the ROC curves [8]. Ultimately, the area under the ROC curve ($A_z$) was used to compare the performance of the logistic regression method as well as an expert radiologist participated in the testing (validating) procedure. To evaluate the performance of an expert radiologist an overall impression of malignancy was ranked using the following five categories scale; benign, probably benign, unsure, probably malignant and malignant. Similarly, to evaluate the performance of the logistic regression classifier, the obtained posterior probability of malignancy i.e. ($\Pr(\Pi_2 | \mathbf{x})$) was classified into the five following categories:

(1) = (0.0-0.2) = " benign"
(2) = (0.2- 0.4) = " Probably benign"
(3) = (0.4-0.6) = " Unsure"
(4) =(0.6-0.8) =" Probably malignant"
(5) = (0.8-1) = "malignant"

**3. Results**

The histologic findings of the biopsies were malignant in 51 cases (42%) and benign in 71 cases (58%). The most common malignant lesions were invasive ductal carcinoma and DCIS, while the most common benign lesions were fibrocystic disease. In our database, we had only 65 cases (53%) with tumor mass in which the size of tumor ranged from 10 mm to 80 mm with a mean of 45mm.

3.1. Radiologist performance

An experienced radiologist read the images and classified them into benign and malignant groups using a five-scale category with increasing likelihood of malignancy. She could not reach to a final diagnosis in 51 cases (41%) and simply classified them as indeterminate or equivocal. The statistical results of sensitivity, specificity and accuracy obtained from the remained cases (n=71) in which the radiologist could reach to a

final decision were 84%, 73% and 76% respectively.

3.2- Logistic regression analysis

First, the estimated logistic regression parameters were obtained from the training sample. Table 2 shows the maximum likelihood estimates of the parameters, standard errors, wald statistic and p-values of the logistic regression model. Taking into consideration all available variables, a logistic regression model established.

The performance of the logistic regression model using the established allocation rule was evaluated. The best performance of the established model was then compared with the reader in terms of accuracy, sensitivity, specificity, false positive fraction, false negative fraction, misclassification rate and correlation with pathology (Table 3). We also applied ROC analysis as a measure of the discriminating ability of a model, with higher areas indicating better predictive ability, to compare the performance of the established model. Using the best results obtained for the model and the radiologist, the ROC analysis was performed (figure 2). The obtained areas under the receiver operating characteristic curves (Az) were presented in table 3.

**4- Discussion**

In this study, we applied an algorithmic model based on the logistic regression analysis to differentiate malignant from benign tumors among a group of 122 patients with approved breast lesions. Our main goal was to investigate weather the used model obtains more reasonable specificity while keeping high sensitivity. Using such a model in clinical practice will lead to decrease the number of cases sent for biopsy; especially in a significant fraction of patients who are going under the biopsy procedure for apparently benign lesions. Using the guidelines for features selection from the previous literatures, the parameters evaluated by a participating radiologist with a high level of experience. The extracted data then presented to the established model. The logit coefficient obtained from wald test in logistic regression model is somehow signifying the importance of any feature in making differentiation between benign and malignant breast tumor. The average output of the logistic regression model yielded a reasonable sensitivity (72%) and accuracy (78%) comparable to the one obtained by the radiologist (84% and 73%). This finding demonstrates a moderate sensitivity with a reasonable specificity for the logistic regression model in differentiating between benign and malignant breast tumors.

Review of the previous studies suggests that the accuracy, sensitivity and specificity of each diagnostic procedure are strongly dependent on the distribution of the benign and malignant patterns among their selected patient's study groups. Therefore, the obtained data by logistic regression models and participated radiologist may not show the exact performance of them. To justify this point, we used ROC analysis to evaluate the performance of model as well as our participated radiologist. By introducing a relative ROC area (Az) of 0.7867 for the logistic regression model compared to 0.7293 obtained by radiologist respectively, the ROC analysis supported and enforced our results.

5- References

[1] Wilson JF, Destouet JM, Winchester DP. 1991 RSNA special focus session: current controversies in the management of ductal carcinoma in situ of the breast. Radiology 1992; 185:77–81.

[2] Parker J, Dance DR, Davies DH, Yeoman LJ, Michell MJ, Humphreys S. Classification of DCIS by image analysis

of calcification from digital mammograms. Br J Radiol 1995; 68:150–9.
[3] Hosmer DW, Lemeshow S. Applied logistic regression, New York: Wiley 1989.
[4]. Ikeda O,Yamashita Y, Morishita S, Kido T, Kitajima M, Okamura K, Fukuda S, Takahashi M. Characterization of breast masses by dynamic enhanced MR imaging. A logistic regression analysis. Acta Radiol. 1999; 40(6): 585-92.
5. Reeves MJ, Osuch JR, Pathak DR. Development of a clinical decision rule for triage of women with palpable breast masses. J Clin Epidemiol. 2003 Jul;56(7):636-45.
6. Cox DR. Some procedures associated with the logistic qualitative response curve. In: David FN. eds. Research papers in statistics: festschrift for J. Neyman. New York: Wiley, 1966; 55-71.
7. Day NE, Kerridge DF. A general maximum likelihood discriminant. Biometrics 1967; 23: 313-23.
8. Metz CE, Some practical issues of experimental design and data analysis in radiological ROC studies, Invest. Radiol. 1989; 24: 234-245.

**Table 1. Evaluated parameters of mammogram of 122 patients, which used as input into the models during the training and validation procedures.**

| Radiological Features | Number of subcategories | code |
|---|---|---|
| Mass Size | 2 | size(mm) |
| Mass shape | 5 | 0 to 4 |
| Mass margin | 4 | 0 to 3 |
| Mass density | 5 | 0 to 4 |
| Asymmetric density | 4 | 0 to 3 |
| Parenchymal distortion | 2 | 0,1 |
| Calcification size | 4 | 0 to 3 |
| Calcification shape | 6 | 0 to 5 |
| Calcification number | 5 | 0 to 4 |
| Calcification density | 3 | 0 to 2 |
| Calcification distribution | 5 | 0 to 4 |
| Associated featuers | 5 | 0 to 4 |

**Table 2. Indicating the maximum likelihood estimates of the parameters, wald statistic and p-values of the logistic regression models fitted to the training sampl**

| Variable | Parameter Estimate | Wald Chi-Square | Pr > Chi-Square |
|---|---|---|---|
| INTERCPT | -0.4539 | 0.1595 | 0.6896 |
| Age | 0.0236 | 1.2653 | 0.2607 |
| Mass size | 0.0115 | 0.5270 | 0.4679 |
| Mass shape | 0.2946 | 1.4367 | 0.1624 |
| Mass margin | 0.2476 | 2.5436 | 0.1107 |
| Mass density | -0.0236 | 1.2653 | 0.2607 |
| Asymmetric density | -0.3921 | 0.5209 | 0.4705 |
| Parenchymal distortion | -0.0235 | 0.0016 | 0.9678 |
| Calcification size | 1.1854 | 1.6111 | 0.2043 |
| Calcification shape | 0.6403 | 2.1826 | 0.0469* |
| Calcification number | 0.1603 | 0.1826 | 0.6692 |
| Calcification density | -0.5769 | 0.5965 | 0.4399 |
| Calcification distribution | 0.6824 | 3.2311 | 0.0392* |
| Associated features | 0.1918 | 0.8521 | 0.3507 |

* significant at level of 0.05

**Table 3. Comparative performance of the participating radiologist, logistic regression model on validation sample (n=40).**

| Parameter | Radiologist | Logistic Discriminant Analysis |
|---|---|---|
| Sensitivity (%) | 84 | 72 |
| Specificity (%) | 73 | 81 |
| Accuracy (%) | 76 | 78 |
| False positive fraction | 27 of 47 | 18 of 22 |
| False negative fraction | 20 of 24 | 13 of 18 |
| Misclassified rate (%) | 34 | 23 |
| $A_z$ * | 0.7293±0.0671 | 0.7867±0.0779 |

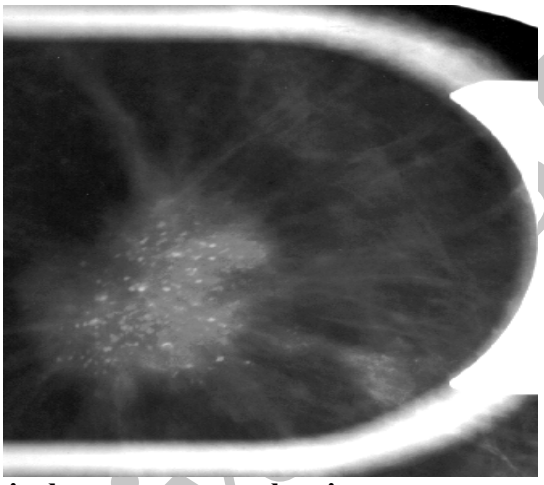

**Figure 1: A typical mammogram showing a mass tumor with cluster of microcalcification**

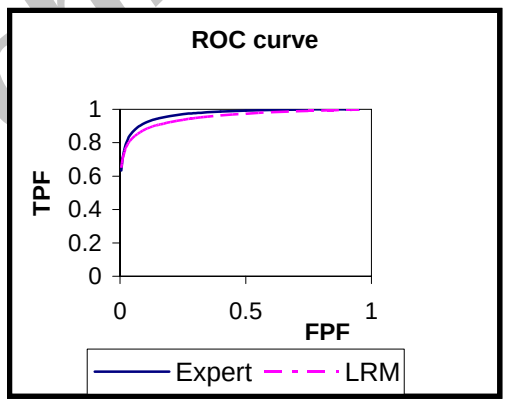


**Figure 2: ROC curves for the expert radiologist and logistic regression model.**