Comparison Multinomial Logistic Regression and Discriminant Analysis in predicting the stage of Breast Cancer

Doungporn Maiprasert[†] **and Krieng Kitbumrungrat**[††],

Rangsit University, Thailand

Summary
This research is a prediction stage of breast cancer group model, the probability that a patient is detected at any stage of breast cancer or non-breast cancer based on the tumor cells with abnormal growth of breast cancer. The independent variable is the tumor cells to grow abnormally: Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8), and Mitoses (X_9). The dependent variable is the probability that the patient is detected at any stage of breast cancer or non-breast cancer based on the tumor cells with abnormal growth of breast cancer by using Ordinal Logistic Regression Model and Discriminant Model. Conclude that Ordinal Logistic Regression Model can use few variables in a prediction stage of breast cancer and Ordinal Logistic Regression Model has classification 55.50% higher than Discriminant Model has classification 54.10%. Ordinal Logistic Regression Model has classification for non-breast cancer patient is 73.60%, breast cancer stage 1 patient is 5%, breast cancer stage 2 patient is 43.6%, breast cancer stage 3 patient is 61.4%. The study results reveal that the Discriminant Analysis can use predicted variables 9 variables. Discriminant Model has classification for non-breast cancer patient is 56%, breast cancer stage 1 patient is 49.10%, breast cancer stage 2 patient is 35.6%, breast cancer stage 3 patient is 72.6% and breast cancer stage 4 patient is 60%.

Key words:
Multinomial Logistic Regression , logistic regression, breast cancer, prediction, Discriminant Analysis

1. Introduction

Cancer is one of the ten causes of the death of the world population. According to the World Health Organization, there were 58 million dead people worldwide in 2008, and 7.6 million of them died of cancer, which is 13% of the total dead people. At present, cancer has caused a great loss of lives of people, specifically liver cancer and breast cancer. Cancer is abnormal tumor cells growth which interfere normal cells, and divide themselves much more quickly than normal cells many times, going beyond the control of a human body. Tumor cells can spread to other parts of a human body, especially to lymph and blood without infecting former tumor cells. The property of tumor cells is they can grow very quickly. The central part of a tumor always lacks of nutrients, resulting in the dead cells of cancer. Consequently, an infection occurs easily because the dead cells and the lymph are good sources of food for diseases, and this can lead to blood infection, which finally causes cancer patients to die. Cancer causes the lack of oxygen in a human body because it consumes a lot of oxygen to help divide its cell, causing white blood corpuscle to work hard so as to eradicate cancer cell, which resulting in the low capability of the human body immune system. Thus, the human body organ in which cancer cells exist will lose its working property, and the spreading of cancer cells can also devastate working property of the nearby organs. The cancer cell will create angiogenesis substance, which causes more blood veins to feed cancerous tumors at a sufficient extent for the growth of cancer cell division. Breast cancer is a disease found mostly in females and it is a genetic transmission disease, causing by abnormal hormone, obesity, food with a high fat, and a spreading of cancer from other part of the organs to the breast.

The 4 ways of treatments of a patient with a breast cancer according to National Cancer Institute, Department of Medical Service, Ministry of Public Health (2008) are operation, radiation therapy, systemic therapy for the first stage cancer, and systemic therapy for the spreading stage cancer. For a medical treatment of a breast cancer, there are many ways of treatment used together to prevent a breast cancer from occurring again, and the consequences of this combination of treatments usually are unexpected side effects, such as nausea, hair falls, diarrhea, and anemia, causing a cancer patient an anxiety and a lot of mental and physical sufferings. A breast cancer needs a vast expense for a treatment, and taking leaves from a job to take care a breast cancer patient causes a cancer patient's relatives to lose a lot of income too. Notwithstanding the complete recovering of a cancer patient in stage 1 from a medical treatment, a cancer patient in a final stage has found unable to completely recover from the condition because cancer is a chronic disease.

Accordingly, the present study paid attention on the study of a prediction of tumor cell with abnormal growth of a breast cancer to detect stages of a breast cancer using 2

Manuscript received April 5, 2012
Manuscript revised April 20, 2012

statistics methods : Multinomial Logistic Regression Analysis and Discriminant Analysis. However, the potential of the Discriminant Analysis rests on ways the data are analyzed, and if an analysis is operated based on the assumption of the model or not. The Discriminant Analysis has a lot of assumptions about the model, that is to say, the interrelationship of groups and predictive variables must be a linear, predictive variables should be distributed together in a dichotomous dimension with not too much interrelationship so as to prevent a dichotomous linear. The Multinomial logistic Regression Analysis has less limitation than Discriminant Analysis.

2. Objectives

The present study aims at studying predictive group discriminant using Multinominal Logistic Regression Analysis and Discriminant Analysis to predict probability of a cancer patient who already had a medical check to determine probability of having a breast cancer, and if detected a breast cancer, the stages of a breast cancer will be predicted based on tumor cells with abnormal growth used as predictive variables to see how predictive results are different and correct employing the two statistics analyses mentioned above.

3. Methodology

The present study is an experimental research conducted with the purpose to study a prediction of tumor cells with abnormal growth and the probability of cancer stages of a breast cancer patient who has medically checked and showed no cancer, or who showed identified cancer stages 1, 2, 3, or 4 using 680 sample database of a medical check of a breast cancer of LopBuri Hospital. The variable (y) was the probability of cancer stages of a breast cancer patient who has medically checked and showed no cancer, or who showed identified cancer stages 1, 2, 3, or 4 and independent variables were Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8) and Mitoses (X_9). The techniques Multinomial Logistic regression Model and Discriminant Analysis were implemented in a prediction of a breast cancer stages in the main study.

3.1.Analysis of Multinomial Logistic Regression

MLR is used as a classification to predict the outcome of biopsy in breast cancer. The MLR is a generalization of the logistic regression model commonly used with the data comprising dependent variables known as “polytomous” and independent variables with numerical or categorical predictors.

The statistical test in MLR includes:

3.1.1 Chi – square is implemented to test these hypotheses:

H_0: The sample has been drawn from population following a specified distribution.

H_1: The sample has not been drawn from population following a specified distribution.

Chi-square test appropriates measures of agreement (or disagreement) between observed and expected frequencies. Chi-square is computed by dividing the squared difference between observed and expected frequencies in each set of frequencies by the expected frequency with the summation of the overall set. The interactions tests were performed to find the significant values of each variable. The significance of the interaction is measured and reported. The test is the cross tabulation test and the values were taken from Pearson Chi-Square.

The Pearson Chi-Square is expressed as

$$X^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

where O_i is observed values

E_i is expected values

X^2 is chi-square value

If the X^2 value is more than the critical value, we reject the null hypothesis.

If the X^2 value is less than the critical value, we accept the null hypothesis.

3.1.2 Maximum likelihood estimate

The principle of maximum likelihood states that the use of estimation of β the value which maximizes the expression in this equation:

$$G^2 = -2[\ln L_p - \ln L_0]; df = p \quad (2)$$

where Lp is likelihood of constant value and group of independence P- value.

L_0 is likelihood of only constant value.

3.1.3 Relationship between independence value and dependence value (Wald test)

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

The Wald test statistic is function of the difference in maximum likelihood estimate (MLE) and the hypothesized

value, normalized by an estimate of the standard deviation of the MLE. The following in equation (3) .

$$\chi^2 = \left[\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)}\right]; df = p \quad (3)$$

where $SE(\hat{\beta}_i)$ is the standard of the maximum likelihood function, estimate is standard of error and df is degree of freedom.

3.1.4 Deviance test (D) is goodness of fit test in MLR in equation 4.

$$D = 2\sum_{i=1}^{n}\sum_{j=1}^{j} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right) \quad (4)$$

where O_{ij} is observed values , E_{ij} is expected values

3.1.5 The simplest optimizing method of discrimination was to maximize to posterior of correct allocation. To obtain the posterior probability the logit coefficients, the following equation is applied:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (5)$$

Where β_0 is the intercept and the β_i i's denotes the unknown logistic regression coefficients of x_i parameters π_i denotes the probability that characteristic will occur. The quantity on the left side of Equation (5) is called a logit. The model can be generalized in the case where the dependent variables, unlike a binary logistic regression model, have more than two categories. Having '4' (stage IV) as the reference category, we can suppose c as the dependent variable with four categories, and the probability of being in category c (c='0' [Benign stage], c='1' [Stage I], c='2' [Stage II] and c='3' [Stage III]) is denoted by P(c) with the chosen reference category, P(4). For such a simple model, MLR with logit link can be represented as

$$\log\left(\frac{P(c)}{P(4)}\right) = \beta_0(c) + \sum_{i=0}^{3} \beta_i(c) x_i, c = 0..3 \quad (6)$$

In this model, the same independent variable appears in each of the c categories, and the separate intercept, $\beta_0(c)$, and slopes (or logit coefficients), $\beta_i(c)$ are usually estimated for selected parameters in each contrast. A way to interpret the effect of independent variables, x_i on the probability of being in category c, is to use predicted probabilities, P(c), for different values of x_i :

$$P(c) = \frac{\exp(\beta_0(c) + \sum_{1}^{n} \beta_i(c) x_i}{1 + \sum_{k=1}^{4} \exp\left(\beta_0(k) + \sum_{i=1}^{n} \beta_i(k) x_i\right)} \quad (7)$$

Then, the probability of being in the reference category, '4' (stage IV), can be calculated by subtraction:

$$P(4) = 1 - \sum_{k=0}^{3} P(k) \quad (8)$$

The category with the highest probability is the final prediction. For detailed descriptions on models with categorical data we refer to (Hayatshahi, S.H.S., 2005).

3.2 Classification of MLR

We wish to classify a patient into one specific class (for example, survival). For many purposes, it will be more helpful to know the predicted probability of survival. A simple but much neglected method is logistic regression which is specified by:

$$P(class2 \mid x) = \frac{e^{\lambda}}{1 + e^{\lambda}} \quad (9)$$

Where $\lambda = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$

$$P(Class1 \mid x) = 1 - P(Class2 \mid x) = \frac{1}{1 + e^{\lambda}} \quad (10)$$

$$\frac{P(class2 \mid x)}{P(Class1 \mid x)} = e^{\lambda} \quad (11)$$

The explanatory variables linearly control the log-odds λ in favour of class 2 (survival). The parameters β are chosen by maximum likelihood that is by maximizing the log-likelihood

$$L = \sum_{i} \log p(class_i \mid x_i) \quad (12)$$

By comparing the patients with features x and the future patients, we will be able to predict P (class 2 | x), probability of survival.

Maximum likelihood is known as 'entropy' fitting and is definitely not common (and supported by amazingly few packages). It is more common to use the regression methods we discuss in section 2, which may be adequate for predicting the class (survival or death) but will be less good for predicting probabilities.

The extension to k > 2 classes is even less well known, although it has a long history. The idea is to take the log-

odds of each class relative to one class, so the model becomes

$$\frac{P(Classj \mid x)}{P(Class1 \mid x)} = e^{\lambda j} \quad , \quad j = 1,2,...,k \tag{13}$$

$$\text{and so} \quad P(classj \mid x) = \frac{e^{\lambda j}}{\sum_c^k e^{\lambda j}} \tag{14}$$

With $\lambda_j = \beta_j^T x$ this is known as MLR. The parameters (β_j) are fitted by maximizing the log-likelihood *L* given in equal (2). There have been surprisingly few non-linear extensions in the statistics literature.

$$P(classj \mid x) = \frac{e^{\lambda j}}{1 + e^{\lambda j}} \quad , \; j = 1,...,k \tag{15}$$

This is an appropriate model for diagnosis where a patient might have none, one or more out of *k* diseases, but not for general classification problems.

3.3 Discriminant Model

Discriminant Function Analysis (DA) undertakes the same task as multiple linear regression by predicting an outcome. However, multiple linear regression is limited to cases where the dependent variable on the Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of weighted combinations of X values. But many interesting variables are categorical.

3.3.1 Fundamental equations for Discriminant Analysis : DA

First, create cross-products matrices for between-group differences and within groups differences, $SS_{total} = SS_{bg} + SS_{wg}$. The determinants are calculated for these matrices and used to calculate a test statistic – either Wilk's Lambda or Pillai's Trace.

Wilk's Lambda follows the equation :

$$\wedge = \left| \frac{S_{wg}}{S_{bg} + S^{wg}} \right| \tag{17}$$

Wilk's Lambda $(\wedge)$ is the ratio of the determinants of the error cross-products matrix to the determinant of the sum of the error and effect cross-products matrices.

The following procedure for calculating approximate F is based on Wilk's Lambda.

$$\text{Approximate } F(df_1 df_{2)} = \left(\frac{1-y}{y} \right) \left(\frac{df_2}{df_1} \right) \tag{18}$$

where df_1 and df_2 are defined below as the degrees of freedom for test the *F* ratio, and y is

$$y = \wedge^{1/s} \tag{19}$$

where p = number of predictor variables.

$\wedge$ is defined in Equation 17.

$$s = \sqrt{\frac{p^2 (df_{effect})^2 - 4}{p^2 + (df_{effect})^2 - 5}} \tag{20}$$

where df_{error} = number of groups times(n-1):
k(n-1).

$$df_1 = p(df_{effect}) \tag{21}$$

$$df_2 = s\left[(df_{error} - \frac{p - df_{effect} + 1}{2} \right] - \left[\frac{p(df_{effect}) - 2}{2} \right] \tag{22}$$

where df_{effect} = number of group minus one (k-1).

The discriminant function score for the i th function is :

$$D_i = d_{i1} Z_1 + d_{i2} Z_2 + ... d_{ip} Z_p \tag{23}$$

Where Z= the score on each predictor, and d_i =discriminant function coefficient. The discriminant function score for a case can be produced with raw scores and unstandardized discriminant function scores. The criminant function coefficients are, by definition, chosen to maximize differences between groups. The mean over all the discriminant function coefficients is zero, with SD equal to one. The mean discriminant function coefficient can be calculated for each group these group means are called Centroids, which are created in the reduced space created by the discriminant function reduced from the initial predictor variables. Differences in the location of these centroids show the dimension along which the group differ. Once the discriminant function are determined groups are differentiated, the utility of these function can be examined via their ability to correctly classify each data point to their a priori groups. Classification functions are derived from the linear discriminant functions to achieve this purpose. Different classification functions are used

and equations exist that are best suited for equal or unequal samples in each group. For cases with an equal sample size for each group classification function coefficient (C_j) is equal to the sum of :

$$C_j = c_{j0} + c_{j1}x_1 + c_{j2}x_2 + ... + c_{jp}x_p \qquad (24)$$

for the j th group, $j = 1,...,k$, x = raw scores of each preditor, c_{j0} = a constant. If W= within group variance – covariance matrix, and M = column matrix of means for group j , then the constant $c_{j0} = (1/2)C_j M_j$.

For unequal sample size in each group :

$$C_j = c_{j0} + \sum_{i=1}^{p} c_{ij}x_i + \ln\left[n_j / N \right] \qquad (25)$$

where n_j = size in group j , N = total sample size.

4. Results

The data for this study was collected from May to September 2008. Data was collected at the Lopburi hospital in Thailand of 680 women.

4.1 Multinomial logistic regression

For experiments, the nine characteristics of breast cancer (Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epitheliai Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses) for input values (x_1-x_9) which each characteristic contains number from 1-10 and five stages of breast cancer.

The statistical test in MLR:

Step1 : An analysis of a breast cancer patient's stages of having a cancer from a medical check and treatment showed that there were not the same ratios of 680 breast cancer patients in each stage of having a breast cancer. For the entire 680 breast cancer patients who had a medical check, it was found that 175 breast cancer patients were not detected a cancer, 29 breast cancer patients were detected a cancer in stage 1, 112 breast cancer patients were detected a cancer in stage 2, 36 breast cancer patients were detected a cancer in stage 3, and 328 breast cancer patients were detected a cancer in stage 4, respectively. When testing the ratios of check and treatment of each stage of a breast cancer, it was found that there was the same ratio of breast cancer patients and breast cancer patients in all stages. (P-value = 0.949 > 0.05 accepted the hypothesis that H_0: the data in each level of dependent variables in each stages of a breast cancer had the same ratios, so the Logit could be used as a link function for the data analysis .

There was a test of the relationship of the two sets of independent variables indicated as tumor cells with abnormal growth and dependent variables indicated as stages of a breast cancer using Logit as a Link Function for the data analysis as shown in Table 1.

Table 1. : Test of a Relationship of Tumor Cells with Abnormal Growth and Stages of A Breast Cancer

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	p-value
Intercept Only	1360.488			
Final	571.876	788.612	36	0.000

According to Table 1, it was found that if the model comprised on constant values, the -2 Log Likelihood would equal 1360.488, and if the model comprised constant values and independent variables indicated as tumor cells with abnormal growth, -2 Log Likelihood would equal 571.876. That was to say, the model with independent variables indicated as tumor cells with abnormal growth was more suitable than the model with mere constant values. To exemplify, at least one independent variable indicated as a tumor cell with abnormal growth had a correlation with a dependent variable indicated as stages of a breast cancer, or a numbers of events of independent variables indicated as tumor cells with abnormal growth had a correlation with situations of not finding a breast cancer, or of finding a breast cancer in any stages. The equation here was (Chi-Square = 788.612 with P-value = 0.000 < 0.05 rejected H_0 independent variables indicated as tumor cell with abnormal growth correlated with dependent variables indicated as stages of a breast cancer).

Step2: There was a test of regression coefficient of the model using Likelihood Ratio Test. The results presented in Table 2.

Table 2 : Result of Likelihood Ratio Test

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2log Likelihood	X^2	df	p-value

Intercept	819.992	248.115	4	.000
Clump Thickness(x_1)	608.852	36.976	4	.000
Uniformity of cell size(x_2)	573.752	1.876	4	.759
Uniformity of cell shape (x_3)	578.574	6.698	4	.153
Marginal Adhesion(x_4)	579.450	7.573	4	.109
Single Epithelial cell size (x_5)	573.045	1.168	4	.883
Bare nuclei (x_6)	615.165	43.289	4	.000
Bland Chromatin x_7)	574.179	2.303	4	.680
Normal nucleoli (x_8)	575.491	3.614	4	.461
Mitoses (x_9)	580.243	8.367	4	.079

*: P-value < 0.05 has a statistical significance at 0.05

According to Table 2, it was found that the Multinomial Logistic Regression Model of tumor cells with abnormal growth had effects on stages of a breast cancer at a P-value of < 0.05, which showed Clump Thickness (X_1) and Bare Nuclei (X_6). Therefore, there was an emergence of a model of accumulated probability prediction for classifying groups of patients based on the health conditions of tumor cells with abnormal growth yielding effects on stages of a breast cancer. The equation derived was

$$\ln\left(\frac{P_i}{1-P_i}\right) = 819.992+608.852X_1+615.165X_6 \quad (26)$$

The likelihood ratio test of tumor cells with abnormal growth with stages of a breast cancer in each stage found that the analysis of the likelihood ratio test of tumor cells with abnormal growth effected each stage of a breast cancer at a P-value <0.05, enabling to create the 4 equations of the Multinomial Logistic Regression Model as follows:

equation with no a breast cancer found

$$\ln\left(\frac{P_0}{1-P_0}\right) = 8.031-0.616x_1-0.46x_3-0.26x_4-0.405x_6 \quad (27)$$

equation of a breast cancer in stage 1

$$\ln\left(\frac{P_1}{1-P_1}\right) = 21.434-0.589x_1-0.464x_6-15.812x_9 \quad (28)$$

equation of a breast cancer in stage 2

$$\ln\left(\frac{P_2}{1-P_2}\right) = 7.574-0.464x_1-0.337x_4-0.432x_6 \quad (29)$$

equation with a breast cancer in stage 3

$$\ln\left(\frac{P_3}{1-P_3}\right) = 6.607-0.55x_1-0.278x_6 \quad (30)$$

For a breast cancer in stage 4, which is a final stage, will depend on the health conditions for the survival. As such, in checking a breast cancer patient and found no cancer, there should have been a consideration on tumor cells with abnormal growth because they might correlate with Clump Thickness (X_1), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), and Bare Nuclei (X_6). For a breast cancer patient with a medical check and found having a breast cancer in stage 1, there should have been a consideration on tumor cells with abnormal growth because they might correlate with Clump Thickness (X_1),, Bare Nuclei (X_6), and Mitoses (X_9)

For a breast cancer patient with a medical check and found having a breast cancer in stage 2, there should have been a consideration on tumor cells with abnormal growth because they might correlate with Clump Thickness (X_1),, Marginal Adhesion (X_4), and Bare Nuclei (X_6).

For a breast cancer patient with a medical check and found having a breast cancer in stage 3, there should have been a consideration on tumor cells with abnormal growth because they might correlate with Bland Chromatin(X_7), and Bare Nuclei (X_6), and a breast cancer in stage 4, which is a final stage, will depend on the health conditions for the survival.

The Multinomial Logistic Regression Model was capable of group classifying 74.1 percent correctly in overall, with a 89.7 percent correct prediction of breast cancer patients with no cancer, 6.9 percent correct prediction of breast cancer patients in stage 1, 12.5 percent correct prediction of breast cancer patients in stage 2, 8.3 percent correct prediction of breast cancer patients in stage 3, and 100 percent correct prediction of breast cancer patients in a final stage.

4.2 Discriminant Analysis

The discriminant analysis to study tumor cells with abnormal growth of breast cancer patients with a medical check and found no cancer, or found having a cancer whether in stage 1, stage 2, stage 3, and stage, with the test to determine if each factor helping classify groups of patients differentiates among the groups using Wilks' Lambda revealed that the diverse factors yielded a statistical significance at a level of 0.05. These factors are 9 tumor cells presented in Table 3.

Table 3 : Statistical Values of Average Equality of Each Factor

Effect	Wilks' Lambda	F-test	p-value
X_1 (Clump Thickness)	0.449	206.917	(0.000)*
X_2 (Uniformity of cell size)	0.439	215.705	(0.000)*
X_3 (Uniformity of cell shape)	0.402	250.589	(0.000)*
X_4(Marginal Adhesion)	0.602	111.607	(0.000)*
X_5(Single Epithelial cell size)	0.625	101.330	(0.000)*
X_6 (Bare nuclei)	0.393	261.113	(0.000)*
X_7 (Bland chromatin)	0.586	119.106	(0.000)*
X_8 (Normal nucleoli)	0.581	121.477	(0.000)*
X_9 (Mitoses)	0.823	36.394	(0.000)*

*: P-value < 0.05

From Table 3, after classifying the factors to identify groups, it was found that the factors for group classification were Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of cell shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin(X_7), Normal Nucleoli (X_8), Mitoses (X_9), which created a model for an analysis of a group classification of breast cancer patients found not having a breast cancer, and found having breast cancer in each stage using Fisher's linear discriminant functions presenting models of groups of breast cancer patients accordingly.

A model of a group of breast cancer patients with no cancer

$$D_0=-3.985 + 0.602X_1 -0.327X_2 +0.054X_3+0.48X_4 +0.478X_5+0.128X_6+0.777X_7-0.019X_8 +0.118X_9 \quad (31)$$

A model of a group of breast cancer patients with a cancer in stage 1

$$D_1=-4.078 + 0.63X_1 - 0.305X_2 +0.021X_3+0.079X_4 +0.50X_5+0.10X_6+0.80X_7 - 0.056X_8 +0.091X_9 \quad (32)$$

which Eigenvalue = 3.816,
Wilks'Lamda = 0.207,
Chi-square = 1058.439,
P-value = 0.000 < 0.05

A model of a group of breast cancer patients with a cancer in stage 2

$$D_2=-3.942+0.583X_1 - 0.319X_2 +0.081X_3+0.037X_4 +0.477X_5+0.123X_6+0.777X_7 - 0.30X_8 + 0.123X_9 \quad (33)$$

which Eigenvalue = 0.002, Wilks'Lamda = 0.997,
Chi-square = 2.127, P-value = 0.000 < 0.05

A model of a group of breast cancer patients with a cancer in stage 3

$$D_3=-4.081 + 0.643X_1 - 0.35X_2 +0.009X_3+0.056X_4 +0.48X_5+0.178X_6+0.733X_7-0.38X_8+ 0.120X_9 \quad (34)$$

which Eigenvalue = 0.001, Wilks'Lamda = 0.999,
Chi-square = 0.664,P-value = 0.004 < 0.05

A model of a group of breast cancer patients with a cancer in stage 4

$$D_4=-17.835+1.482X_1 - 1.77X_2 +0.453X_3+0.265X_4 +0.68X_5+0.985X_6+0.955X_7 + 0.254X_8 +0.20X_9 \quad (35)$$

which Eigenvalue = 0.000, Wilks'Lamda = 1.000, Chi-square = 0.012, P-value = 1.000 > 0.05, showing that a D_4 model did not correlated with the model because a breast cancer in stage 4 is a final stage and depends on the health conditions for the survival.

4.3 Results of the Efficiency Comparison

The efficiency of the Discriminant analysis and the Multinomial Logistic Regression Analysis considered based on the correct rations of classification was presented in Table 4.

Table 4 The number and percentile of breast cancer patients in correct groups of classification based on Discriminant analysis and the Multinomial Logistic Regression Analysis employed with 683 samples of breast cancer patients

Table 4 : Comparison between the Discriminant analysis and the Multinomial Logistic Regression Analysis

Model	Discriminant	Multinomial Logistic Regression
Predictive variables		
-Benign Stage	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_1, X_3, X_4, X_6
-Stage 1	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_1, X_6, X_9
-Stage 2	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_1, X_4, X_6
-Stage 3	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	X_1, X_6
-Stage 4	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	Stage4 with referenced variables
Capability of group classification (%)		
-Benign Stage	6.0%	89.7%
-Stage 1	37.9%	6.9%
-Stage 2	45.5%	12.5%
-Stage 3	25.0%	8.3%
-Stage 4	92.7%	100%
Total	55.3%	74.1%

As presented in Table 4, it was found that the discriminant analysis used all variables to predict all stages of having a breast cancer, or not having a breast cancer with a capability of 55.30 percent correctly predicting in overall, 6 percent correctly predicting breast cancer patients having no a cancer, 37.90 percent correctly predicting breast cancer patients in stage 1, 45.5 percent correctly predicting breast cancer patients in stage 2, 25 percent correctly predicting breast cancer patients in stage 3, and 92.7 percent correctly predicting breast cancer patients in stage 4. In contrast, the Multinomial Logistic Regression used some variables to predict all stages of having a breast cancer, which its capability of correctly predicting was higher than the Discriminant Model. The Multinomial Logistic Regression was capable of 74.1 percent correctly classifying groups in overall, 89.7 percent correctly predicting breast cancer patients having no cancer, 6.9 percent correctly predicting breast cancer patients in stage 1, 12.5 percent correctly predicting breast cancer patients in stage 2, 8.3 percent correctly predicting breast cancer in stage 3, and 100 percent correctly predicting breast cancer patients in stage 4.

5. Conclusions and Discussion

The prediction of the probability for the classification of breast cancer patients with having no cancer or having cancer in any stages was considered an essential tool to create health conditions of tumor cells with abnormal growth. In the present study, there were 9 types of tumor cells with abnormal growth indicated by 2 statistical methods of prediction: Multinomial Logistic Regression Analysis and Discriminant Analysis. The results of an analysis revealed that the Multinomial Logistic Regression Analysis was capable of 55.50 percent correctly predicting in overall, which was more correct than the analysis done by Discriminant Analysis, giving a 54.10 percent correct prediction in overall. The Multinomial Logistic Regression Analysis was capable of correctly classifying 73.60 percent for breast cancer patients with no cancer, 5 percent for breast cancer patients with stage 1, 43.6 percent for breast cancer patients with stage 2, 43.6 percent for breast cancer patients with stage 3, 61.4 percent for breast cancer patients with stage 4, and 76.40 percent for assigning predictive variables to the model indicating that breast cancer patients with no cancer correlated with Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), (Bland Chromatin(X_7), Normal Nucleoli (X_8), and Normal Nucleoli (X_8), the breast cancer patients with stage 1 correlated with Clump Thickness (X_1), and Bare Nuclei (X_6), the breast cancer patients with stage 2 correlated with Clump Thickness (X_1), Uniformity of Cell Shape (X_4), Bare Nuclei (X_6), and Normal Nucleoli (X_8), the breast cancer patients with stage 3 correlated with (Bland Chromatin(X_7), Normal Nucleoli (X_8), and Mitoses (X_9), and the breast cancer patients with final stage, which was a referenced stage using health conditions for a survival.

The Discriminant Analysis used all the 9 predictive variables. The Discriminant Analysis used all the 9 variables to predict stages of breast cancer patients, who had a breast cancer, or who did not have a breast cancer. Its capability of correctly classifying groups of breast cancer patients in overall was 55.3 percent, breast cancer patients in stage 1 was 6 percent, breast cancer patients in stage 2 was 37.90 percent, breast cancer patients in stage 3 was 45.5 percent, breast cancer patients in stage 3 was 25 percent, and breast cancer patients in a final stage was 92.7 percent. The models of factors for classifying groups was Fisher's linear discriminant functions indicating the following groups of breast cancer patients.

A model of a group of breast cancer patients having no cancer was:

$$D_0 = -3.985 + 0.602X_1 - 0.327X_2 + 0.054X_3 + 0.48X_4 + 0.478X_5 + 0.128X_6 + 0.777X_7 - 0.019X_8 + 0.118X_9$$

A model of a group of breast cancer patients having a breast cancer in stage 1 was:

$$D_1 = -4.078 + 0.63X_1 - 0.305X_2 + 0.021X_3 + 0.079X_4 + 0.50X_5 + 0.10X_6 + 0.80X_7 - 0.056X_8 + 0.091X_9$$

A model of a group of breast cancer patients having a breast cancer in stage 2 was:

$$D_2=-3.942+0.583X_1 - 0.319X_2 +0.081X_3+0.037X_4 +0.477X_5+0.123X_6+0.777X_7 - 0.30X_8 + 0.123X_9$$

A model of a group of breast cancer patients having a breast cancer in stage 3 was:

$$D_3=-4.081 + 0.643X_1 - 0.35X_2 +0.009X_3+0.056X_4 +0.48X_5+0.178X_6+0.733X_7 - 0.38X_8 + 0.120X_9$$

A model of a group of breast cancer patients having a breast cancer in stage 4 was:

$$D_4=-17.835+1.482X_1 - 1.77X_2 +0.453X_3+0.265X_4 +0.68X_5+0.985X_6+0.955X_7+0.254X_8 +0.20X_9,$$

indicating that D_4 model did not correlated with the model because a breast cancer in stage 4 is a final stage and depends on the health conditions for the survival.
Therefore, the models for statistical classification would be effective in case the data used for an analysis had distributions in accordance with the assumptions of statistical paramatrix of groups classification used for predictions in order to enhance health conditions of the breast cancer patients. These 2 statistical predictions were Multinomial Logistic Regression Analysis and Discriminant Analysis, which later should be operated with a horizontal paramatrix for health conditions of patients, for example neutral networks, decision trees, and trait recognition approach.

Acknowledgements

The author would like to express gratitude to Associate Professor Dr. Chom Kimpan for permission of the extension of this research.

References

[1] Agreti, A. (2002). *Categorical Data Analysis*, (2nd ed.). New York: John Wiley&Sons.
[2] Bandhita P., Noparat T.(2006, June). *Ordinal Regression Analysis in factors related to Sensorial Hearing Loss of the Employee in Industrial factory in Lampang Thailand.* Mathematic, Statistics and Their Application, Penang.
[3] Hayatshahi, S.H.S., Abdolmaleki, A., Safarian, S. & Khajeh, K.(2005,October). *Non-linear quantitative structure-activity relationship foradenine derivatives as competitive inhibitors of adenosine deaminase.* Biochem Biophys Research Comunication., 338, 1137-1142. Retrieved from http://elsevier.com/locate/ybbrc
[4] Hosmer, D.W.,&Lemeshow, S.(2000). *Application logistic regression.*New York : John Wiley & Sons.
[5] Jobson , J.D. (1992). *Applied multivariate data analysis*. (2 nd ed.). Springer, New York : Berlin Heidelberg.
[6] Klecka, William R. (1980). *Discriminant Analysis.* Quantitative Applications in the Social Sciences Series, No.19. Thousand Oaks, CA:Sage Publications.
[7] Lachenbruch, P. A. (1975). *Discriminant Analysis*. NY:Hafner. For detailed notes on computations.
[8] Overall, J.E. and C.J.Klett. (1972). *Applied Multivariate Analysis.* McGraw-Hill: New York.
[9] Stephenson, B. (2008). Binary response and logistic regression analysis. *Retrieved from www.public.iastate.edu / ~stat415/stephenson/stat415_chapter3.pdf.*
[10] Swets, J. (1988). *Measuring the accuracy of diagnostic systems*. Science, 240, 1285-1293. doi : 101126/science.3287615
[11] Swets, J.A., Dawes, R.M., Monahan, J. 2000, October). Better decisions through science. Scientific American, 283, 82-87.
[12] VECCHIO, T (1966). *Predictive value of a single diagnostic test in unselected populations*. New England Journal of Medicine, 274, 1171-1173.
[13] Wingo P.A., Tong, T., Bolden, S. (1995). *Cancer statistics*, CA Cancer J Clin, 46(1), 5-27.doi : 10.3322/canjclin.46.1.5

Doungporn Maiprasert was born in Lopburi,Thailand, in 1974. She received the B.S. degree in computer science from Thepsatri Rajabhat University. Thailand in 1995 and received the M.S. degree in Information technology from Rangsit University. Thailand in 2005. Her research interests include artificial intelligence and the application of the statistical.

Krieng Kitbumrungrat was born in Bangkok, Thailand, in 1964. He received the B.S. degree in Statistics from Nation Institute of Development Administration (NIDA) University. Thailand in 1991. He received Doctor of Philosophy in Mathematics from Mahidol University.Thailand in 2004. His research interests include mathematics and the application of the statistical.