Paper 1739-2014

Medical Scoring for Breast Cancer Recurrence

Nurul Husna Jamian, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
Yap Bee Wah, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
Nor Aina Emran, Department of Surgery, Hospital Kuala Lumpur, 50586 Kuala Lumpur, Malaysia.

ABSTRACT

Breast cancer is the most common cancer among females globally. After being diagnosed and treated for breast cancer, patients fear the recurrence of breast cancer. Breast cancer recurrence (BCR) can be defined as the return of breast cancer after primary treatment, and it can recur within the first three to five years. BCR studies have been conducted mostly in developed countries such as the United States, Japan, and Canada. Thus, the primary aim of this study is to investigate the feasibility of building a medical scorecard to assess the risk of BCR among Malaysian women. The medical scorecard was developed using data from 454 out of 1,149 patients who were diagnosed and underwent treatment at the Department of Surgery, Hospital Kuala Lumpur from 2006 until 2011. The outcome variable is a binary variable with two values: 1 (recurrence) and 0 (remission). Based on the availability of data, only 13 categorical predictors were identified and used in this study. The predictive performance of the Breast Cancer Recurrence scorecard (BCR scorecard) model was compared to the standard logistic regression (LR) model. Both the BCR scorecard and LR model were developed using SAS® Enterprise Miner™ 7.1. From this exploratory study, although the BCR scorecard model has better predictive ability with a lower misclassification rate (18%) compared to the logistic regression model (23%), the sensitivity of the BCR scorecard model is still low, possibly due to the small sample size and small number of risk factors. Five important risk factors were identified: histological type, race, stage, tumor size, and vascular invasion in predicting recurrence status.

INTRODUCTION

Generally, cancer begins when abnormal cells in a part of the body start to overgrowth and invade other tissues while breast cancer is a malignant (cancer) tumour that starts from cells of breast [1]. Breast cancer is the commonest diagnosed cancer with a substantial high proportion among female globally. In addition breast cancer is 100 times more common in women than in men since women have more estrogens and progesterone hormones which promote breast cancer cell growth [1].

In United States, breast cancer is the second leading cause of cancer death for women after lung cancer and it was reported that approximately 39 510 women and 410 men will die from the disease in 2012 [2]. It found that a woman has a 1 in 8 chance of developing invasive breast cancer during her lifetime which about 1 in 11 in 1975 [3]. There are some factors that may increase a woman's risk of breast cancer such older age, genetic factors, family history of breast or ovarian cancer, long menstrual history, nulliparity, older than 30 years of age at first full-term pregnancy, daily alcohol consumption, use of combined postmenopausal hormone replacement therapy (HRT), postmenopausal, obesity and ionizing radiation [2].

In Asia, the incidence of breast cancer is lower than in the West but it is increasing rapidly while Malaysia has the higher incidence rate as compared to the neighbouring countries especially Indonesia and Thailand [4]. Recent statistics in Malaysia reported a total of 18 219 cancer cases were diagnosed in 2007 and registered at the National Cancer Registry (NCR) and the five most common cancers among females were breast, colorectal, cervix, ovary and lung. There were 3242 female breast cancer cases registered in NCR for that year, accounted for 32.1% of all cancer cases registered [5]. The most common cancer occurred among women in Malaysia is breast cancer (26.5%). It is the highest number of cases diagnosed compared to other cancers such as cervix uterine (12.6%), colorectal (9.9%), lung (5.8%) and ovary (5.4%) [6].

Customarily, most breast cancer patients undergo treatment once they were diagnosed with breast cancer of various stages. Many of these patients have high chances of survival with good medical treatment and lifestyle. Unfortunately, there are breast cancer patients who experience a recurrence or the development of a second primary cancer. Recurrence can be defined as the return of breast cancer after primary treatment that can recur anytime usually within the first three to five years [7]. In Malaysia, The Preliminary Report in 2008 by National Cancer Patient Registry (NCPR) consists of reports on follow-up breast cancer patients from NCPR-Breast Cancer. There were 108 out of 154 breast cancer patients with available follow-up information within the period of 1st June 2008 to 31st December 2008. The status of patients in this follow-up period showed that 94.4% patients are disease free while about 5.6% patients are recurrent cases [8].

There is still lack of studies on the risk factors of breast cancer recurrence in Malaysia where most of the studies focus on survival rate and the risk factors of breast cancer [9, 10, 11]. However, breast cancer recurrence studies have been mostly conducted in developing countries such United States, Japan and Canada. Thus, the main objective of this study is to develop a breast cancer recurrence (BCR) scorecard model using Malaysian breast cancer data and compare it with Logistic Regression model using SAS® Enterprise Miner™ 7.1.

LITERATURE REVIEW

A. Risk Factors

There is still lack of studies on the risk factors of breast cancer recurrence in Malaysia. Most studies focus on survival rate of breast cancer and the risk factors [9-11]. Breast cancer recurrence studies have been mostly conducted in developing countries such as the United States, Japan and Canada. The selection of variables to be included in this study was based on a review of the previous studies on the risk factors of breast cancer.

Age is found to be the most important risk factor for developing breast cancer. The lifetime risk of developing breast cancer is 1 in 8 for a woman who lives to be 90 [12]. This means that 7 out of 8 women will not develop breast cancer in their lifetime while the risk for younger women is much lower. In Malaysia, the most common age group diagnosed with breast cancer is 50.6% in women between 40 to 49 years, 16.8% in women below 40 years and 2% in women below 30 years [13-15]. Women below age 35 years who had breast conserving therapy have 9.24 times higher risk of recurrence than women over 60. Meanwhile for radical mastectomy, there was no difference in recurrence observed among the different age groups [16].

In the United States, the occurrence of breast cancer is more common among White women than Latina, Asian, or African American women [17]. In Malaysia, the highest number of deaths among breast cancer patients is Malay women compared to other ethnicity since Malay women are often diagnosed at later stages and with larger tumours [13-14]. In studies on breast cancer in relation to the three races in Malaysia there are significant difference among Malay, Chinese and Indian with the most common in the Chinese followed by Indians and then, Malays. [18-19].

A small number of women have an increased risk of developing breast cancer because they have a significant family history [12]. Therefore, it is essential to know the number of relatives with breast cancer, how many close relatives (first degree relatives) had breast cancer and the age that the breast cancer was diagnosed [20]. According to [21], family history is a potential risk factor of breast cancer recurrence.

Stage of disease is a priority for women with breast cancer [22]. In Malaysia, about 40% Malays are diagnosed at later stages (Stage 3 and 4) compared to only 15% of Chinese and 20% of Indians [13]. According to [4], Asian women are diagnosed at later stages of the disease, hence the mortality rate is higher and the patients detected at stages 3 to 4 is about 50-60% at HKL and 30-40% at UMMC. The tumour size in the detection phase of breast cancer patients is used to classify the stages level and to decide the best treatment to be employed [24]. The stages of breast cancer also have significant effect on breast cancer recurrence [21, 23, 25-28].

A positive receptor status indicates that the tumour will likely respond to hormonal therapy [29]. Hormone receptor status reflects whether the tumour is estrogens receptor positive (ER+) or not (ER-), progesterone receptor positive (PgR+) or not (PgR-). About two-third of all breast cancer contain significant levels of estrogens receptor positive, ER+. However, ER+ tumours tend to grow less aggressively and may respond favourably to treatment with hormones. This status may have some prognostic information and it used to plan treatment [24]. Therefore, Hormone Receptor Status (ER/PR) is a crucial risk factor of breast cancer recurrence [23, 26, 30].

Lymph node involvement is the most important prognostic factor with 10 year survival rate decreasing from 75% with negative lymph nodes to just 25-30% in lymph node-positive patients [31]. Women who have lymph node involvement are more likely to have a recurrence. However, the concern is whether the tumour has spread to the lymph nodes at the time of diagnosis (node positive) and the number of lymph nodes in which the cancer has been found. [24]. Recent study showed that lymph nodes status has significant effect on breast cancer recurrence [21, 26, 28].

B. Credit Scoring

This study used the method of credit scoring in business applications to develop a medical scorecard for prediction of breast cancer recurrence. Credit scoring is often used to predict the credit risk or probability that a loan applicant or existing borrower will default or become delinquent. Besides, credit scoring means applying a statistical model to assign a risk score to a credit application or to an existing credit account from historical data [32-36].

Credit scoring uses quantitative measures of the performance and characteristics of past loans to predict the future performance of loans with similar characteristics [33]. Many banks used scoring to evaluate small-business

loan applications. The method produces a "score" that a bank can use to rank its loan applicants in terms of risk. Thus, developers need to analyze historical data on the performance of previously made loans to build a scoring model or "scorecard". In most scoring systems, a higher score indicates lower risk, and a lender sets a cut-off score based on the amount of risk it is willing to accept. The lender would approve applicants with scores above the cut-off and deny applicants with scores below [32]. However, the application of credit scoring has grown from making credit decisions related to housing, insurance, basic utility services, club fee payment and even employment [35-36]. Credit scoring greatly reduces the time of loan approval process where the time savings means cost savings to the bank and benefits the customer as well [32]. Credit scoring can be performed quickly and efficiently obtain fact-based and accurate predictions of the credit risk of individual applicants [34].

METHODOLOGY

This study was registered with the National Medical Research Register (NMRR). Institutional Approval of Clinical Research Centre (CRC) and Ethics Approval of Medical Research and Ethics Committee (MREC) were obtained prior to data collection. There were 1149 breast cancer patients who were diagnosed and underwent treatment at Department of Surgery in Hospital Kuala Lumpur from 2006 until 2011. After data cleaning, the sample consists of 454 cases for further analysis. Based on literature review and availability of data, 13 variables were included as predictors of breast cancer recurrence status. The description of variables and the frequency distribution is shown in Table 1.

Variables	Description	Role	No. of Patient	Percentage (%)
Recurrence status	1: Recurrence 0: Remission	Target	72 382	15.9 84.1
Age group	1: Less than 40 years 2: 40 – 59 years 3:More than 59 years	Input	49 277 128	10.8 61.0 28.2
Marital status	1: Single 2: Married	Input	64 390	14.1 85.9
Race	1: Malay 2: Chinese 3: Indian	Input	271 111 72	59.7 24.4 15.9
Family history	1: Yes 0: No	Input	57 397	12.6 87.4
Stage	1: I 2: II 3: III 4: IV	Input	42 243 140 29	9.3 53.5 30.8 6.4
Tumour size	1: Less than 3 cm 2: 3cm-5.9cm 3: 6 cm and above	Input	151 229 74	33.3 50.4 16.3
Histological grade	1: Grade I 2: Grade II 3: Grade III	Input	87 216 151	19.2 47.6 33.3
Histological type	1: Ductal infiltrating 2: Lobular infiltrating	Input	277 177	61.0 39.0
Lymph nodes status	1: 0 2: 1-3 3: 4-9 4: 10 and above	Input	227 85 74 68	50.0 18.7 16.3 15.0
Vascular invasion	1: Present 0: Absent	Input	146 308	32.2 67.8
Estrogens receptors	1: Positive 2: Negative	Input	280 174	61.7 38.3
Progesterone receptors	1: Positive 2: Negative	Input	250 204	55.1 44.9
CerB2	1: Positive 2: Negative	Input	235 219	51.8 48.2

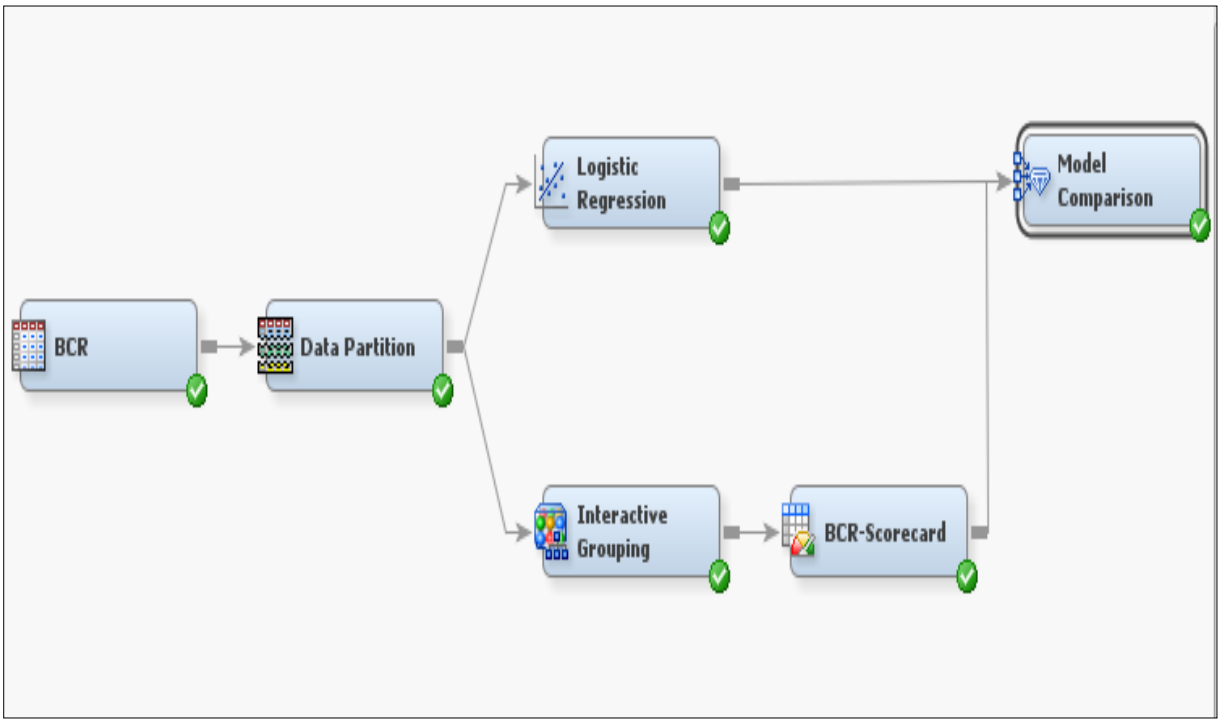
Table 1. Description of Variable and Frequency Distribution

To build the models, the sample of 454 patients was first partitioned into two sub samples at ratio 70:30 for training and validation respectively. The training data used for model fitting while validation data used for model validation. The sample size for the training and validation samples is depicted in Table 2.

Partition	Recurrence	Remission	Total
Training (70%)	49	267	316
Validation (30%)	23	115	138
Total	72	382	454

Table 2. The Partitioned Data for Training and Validation

The scorecard was developed using SAS® Enterprise Miner™ 7.1. The process flow diagram is shown in Display 1. The dataset node (BCR node) firstly dragged into the diagram workspace as the data source of the study. Then the Data Partition node was connected to BCR node in order to split the original dataset into training and validation dataset. Thereafter, the Logistic Regression node and Interactive Grouping were connected to the Data Partition node. The BCR-scorecard requires the Interactive Grouping node and Scorecard node. Then the performances for both models were assessed and compared by connecting them towards Models Comparison node.



Display 1. Flow Diagram of Developing Medical Scoring Models.

A. BCR-Scorecard Model

Basically, the procedures for constructing the BCR-Scorecard model are classing and score points scaling. Classing is the process of automatically and/or interactively binning and grouping input that takes place in the Interactive Grouping node in order to manage the number of attributes per characteristics, improve the predictive power of the characteristics, select predictive characteristics and obtain the Weight of Evidence (WOE). WOE is used to class and select the characteristics based on the score points assigned for each attribute in order to determine the relative risk of the attributes. A high negative WOE implies the high risk (Recurrence) of an attribute while a high positive implies the low risk (Remission). The formula of WOE is as follows:

$$\text{WOE attribute} = \ln\left(\frac{P_{goodattribute}}{P_{badattribute}}\right)$$

where

$$P_{goodattribute} = \frac{\#\ goods_{attribute}}{\#\ goods}$$

$$P_{badattribute} = \frac{\#\ bads_{attribute}}{\#\ bads} \qquad (1)$$

Then, the predictive powers of each attributes are assessed by their Information Value (IV). IV is the weighted sum of the WOE of the attributes that is weighted by the difference between the proportion of “good” and “bad” in the respective attribute. It is able to separate high risks (Recurrence) from low risks (Remission) in order to select

the characteristics for inclusion in the scorecard model. If the IV is greater than 0.10, the variable is chosen as an input variable. The formula of IV is as follows:

$$\text{IV}= \sum(\text{P}_{\text{goodattribute}} - \text{P}_{\text{badattribute}}) * \text{WOE} \quad (2)$$

For the score-points scaling process, the WOE and the regression coefficient of its characteristic will be multiplied to obtain score points for each attribute. The score points for each attribute is calculated as follows:

$$\text{Score}= -\left(\text{WOE} * \beta_i + \frac{\alpha}{n}\right) * \text{factor} + \frac{\text{offset}}{n}$$

where

β_i = regression coefficient for WOE
α = estimate of intercept
n = number of variables selected
Factor = points to double odds/log 2
Offset=(score-factor)*log(odds) (3)

The cut-off score is obtained from the value of Kolmogorov Smirnov statistic or K-S test result. The selected of cut-off score is based on the highest K-S test where the K-S test measures the distance between the distribution functions of the "good" and "bad".

B. Logistic Regression Model

Logistic regression model is an appropriate model when the dependent variable is dichotomous (binary). The logistic regression model used to determine the probability of breast cancer recurrence P(Y=1), that is:

$$P(Y_i = 1) = \frac{1}{1+e^{z_i}}$$

where

$$z_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

then,

P (Y_i=1)= the probability of event of interest, for case i, i=1,2,...,n
β_0= the constant of the equation.
β_j= the coefficient of the predictor variable, X.
X_j = the predictor variable, j=1,2,...,k (4)

The Wald statistic is commonly used to test the significance of the individual logistic regression coefficient resulting in identifying the variables that influence Breast Cancer Recurrence.
H_0: $\beta_i = 0$ (the independent variable has no effect on Breast Cancer Recurrence).
H_i: $\beta_i \neq 0$ (the independent variable has an effect on Breast Cancer Recurrence).

A Wald test calculates a z statistic, z = $(\frac{B}{SE})$ where B is the estimated regression coefficient and SE is the estimate of the standard error. The Z is then squared and the Wald statistic follows the Chi square distribution ($z^2 \sim \chi_1^2$):

$$\text{Wald Statistic} = \left(\frac{\text{B}}{\text{SE}}\right)^2 \quad (5)$$

Reject H_0 if Wald statistic > $\chi^2_{\alpha,1}$ or p-value < 0.05

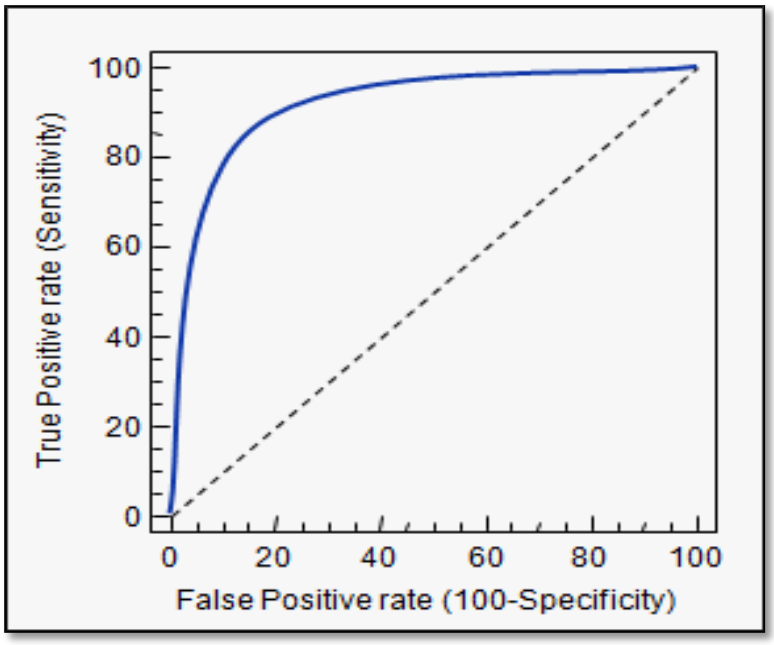
Odds ratio is measure of association in term of how much each significant independent variable is more or less likely to cause the outcome of interest (Y=1). In this case Y=1 is the 'Recurrence' event of interest. The Odds ratio is obtained as follows:

$$\text{Odds ratio} = e^{\hat{\beta}_1}$$

where $\hat{\beta}_1$ the estimated coefficient (6)

C. MODELS COMPARISON

The performance of the BCR-scorecard model and logistic regression model were assessed based on the Receiver Operating Characteristic (ROC) Curve, misclassification rate, sensitivity and specificity. For Receiver Operating Characteristic (ROC) Curve, the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test as illustrated in Display 2.



Display 2. Receiver Operating Characteristic (ROC) Curve
(Source: Zweig and Campbell, 1993)

Based on the Confusion matrix in Table 3, True Positive is the number of patients who are predicted as recurrence among the recurrence patients. While True Negative the number of patients predicted as remission among the remission patients. False Positive indicates patients predicted as remission among recurrence patients. False negative is when patients are predicted as recurrence among remission patients. Misclassification rate is percentage of patients who were misclassified. Accuracy rate indicates how well the model correctly predicted the outcome of recurrence and remission. Sensitivity is the probability that the model can correctly predict recurrence while specificity is the probability of the model can correctly predict remission. A good model consists of lower misclassification rate, higher accuracy rate and high sensitivity. The formula of misclassification rate, accuracy, sensitivity and specificity are shown below:

$$\text{Misclassification} = \frac{FP+FN}{TP+FP+FN+TN} \quad (7)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (10)$$

		Actual		Total Predicted
		Recurrence	Remission	
Predicted	Recurrence	True Positive (TP)	False Positive (FP)	TP+FP
	Remission	False Negative (FN)	True Negative (TN)	FN+TN
Total Actual		TP+FN	FP+TN	TP+FP+FN+TN

Table 3. Confusion Table

RESULTS

A. BCR-Scorecard Model Results

The Interactive Grouping node (IGN) is employed for grouping the variables. The Screenshot of Output Variables is shown in Display 3.

Results - Node: Interactive Grouping Diagram: ANALYSIS BCR

File Edit View Window

Output Variables

Variable	Gini Statistic	Information Value	Level for Interactive	Exported Role	New Role	Pre-Defined Grouping	Level	Label	Information Value Ordering
Stage	17.71	0.222	NOMINAL	Input	Input		NOMINAL	Stage	1
Histo_type	18	0.154	BINARY	Input	Input		BINARY	Histo_type	2
Vascular	18.474	0.148	BINARY	Input	Input		BINARY	Vascular	3
Race	18.352	0.143	NOMINAL	Input	Input		NOMINAL	Race	4
Tumour_sz	18.299	0.126	NOMINAL	Input	Input		NOMINAL	Tumour_sz	5
Age_group	12.161	0.083	NOMINAL	Rejected	Rejected		NOMINAL	Age_group	6
LN_status	13.315	0.067	NOMINAL	Rejected	Rejected		NOMINAL	LN_status	7
Marital_status	8.049	0.048	BINARY	Rejected	Rejected		BINARY	Marital_status	8
PR	10.663	0.046	BINARY	Rejected	Rejected		BINARY	PR	9
Family_hist...	5.32	0.03	BINARY	Rejected	Rejected		BINARY	Family_hist...	10
CerB2	7.873	0.025	BINARY	Rejected	Rejected		BINARY	CerB2	11
Histo_Grade	5.74	0.012	NOMINAL	Rejected	Rejected		NOMINAL	Histo_Grade	12
ER	5.006	0.011	BINARY	Rejected	Rejected		BINARY	ER	13

Display 3. Screenshot of Output Variables

Display 4 illustrates the Screenshot for Kolmogorov Smirnov Table that demonstrates the cut-off score in this study was 140 due to the highest K-S statistic, 0.418023. It means that 140 is the minimum acceptable level of risk. Hence, patients who score above 140 are less likely to suffer recurrence while those who score below 140 have higher chances of recurrence.

Results - Node: BCR-Scorecard Diagram: ANALYSIS BCR

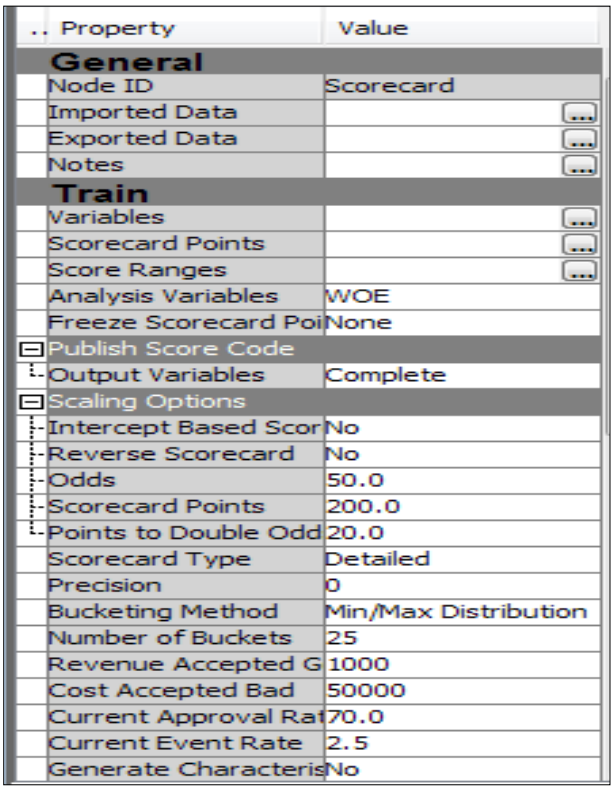
File Edit View Window

Table: Kolmogorov-Smirnov Plot

Bucket	event	Data Role	Kolmogorov-Smirnov Statistic ▼	Score Bucket	...	Cutoff Score	Population Percentage	Type	Frequency	Average Scorecard Points	Empirical Odds
111		TRAIN	0.418023	133 <= Score < 140		140	59.81013	0	38	137.6842	-2.89037
101		TRAIN	0.408851	126 <= Score < 133	...	133	76.58228	0	53	129.9623	-0.92954
121		TRAIN	0.324008	140 <= Score < 147		147	47.78481	0	8	144.125	-2.83321
131		TRAIN	0.314454	147 <= Score < 154		154	45.25316	0	27	148.4444	-2.07944
91		TRAIN	0.304976	119 <= Score < 126		126	86.07595	0	30	121.4333	-2.19722
81		TRAIN	0.288695	112 <= Score < 119		119	91.13924	0	16	115.3125	-0.51083
141		TRAIN	0.2783	154 <= Score < 161		161	36.70886	0	27	155.037	-2.52573
111		VALID	0.26087	133 <= Score < 140		140	52.89855	0	14	136.9286	-1.79176
101		VALID	0.226087	126 <= Score < 133	...	133	64.49275	0	16	130.1875	-1.09861
151		TRAIN	0.212566	161 <= Score < 168		168	28.16456	0	29	163.4828	-2.60269
81		VALID	0.2112	<= Score < 119		119	83.33333	0	16	114.5	-1.09861
121		VALID	0.2140	<= Score < 147		147	42.75362	0	2	146	-1.60944
71		TRAIN	0.192464	105 <= Score < 112	...	112	94.3038	0	10	108.5	-0.40547
141		VALID	0.191304	154 <= Score < 161		161	33.33333	0	7	154.8571	-2.70805
91		VALID	0.182609	119 <= Score < 126		126	71.73913	0	10	121.5	-2.19722
131		VALID	0.182609	147 <= Score < 154		154	41.30435	0	11	149.3636	-1.50408
151		VALID	0.165217	161 <= Score < 168		168	28.26087	0	18	163.7222	-2.07944
161		TRAIN	0.152259	168 <= Score < 175		175	18.98734	0	21	171.5714	-2.99573
61		TRAIN	0.13330398	<= Score < 105		105	95.88608	0	5	102.6	-0.40547
61		VALID	0.11304398	<= Score < 105		105	92.75362	0	11	102.2727	-1.50408
171		TRAIN	0.09776	175 <= Score < 182		182	12.34177	0	9	179.5556	-2.07944
181		TRAIN	0.088206	182 <= Score < 189		189	9.493671	0	6	187.1667	-2.56495
191		TRAIN	0.086142	189 <= Score < 196		196	7.594937	0	4	192.75	-1.09861
161		VALID	0.078261	168 <= Score < 175		175	15.21739	0	4	171.5	-2.19722
71		VALID	0.078261	105 <= Score < 112		112	84.78261	0	2	106	0
201		TRAIN	0.074906	196 <= Score < 203		203	6.329114	0	9	201	-2.94444
41		TRAIN	0.06665	184 <= Score < 91		91	98.41772	0	4	87.25	-1.09861
51		TRAIN	0.06290691	<= Score < 98		98	97.1519	0	4	94	1.098612
51		VALID	0.05217491	<= Score < 98		98	95.65217	0	4	94	-1.09861
201		VALID	0.052174	196 <= Score < 203		203	4.347826	0	1	201	-1.09861
211		VALID	0.043478	203 <= Score < 210		210	3.623188	0	2	207.5	-1.60944
171		VALID	0.043478	175 <= Score < 182		182	12.31884	0	8	180	-1.94591
211		TRAIN	0.041199	203 <= Score < 210		210	3.481013	0	3	207	-1.94591
31		TRAIN	0.03707177	<= Score < 84		84	99.05063	0	2	78	0

Display 4. Screenshot for Kolmogorov Smirnov Table

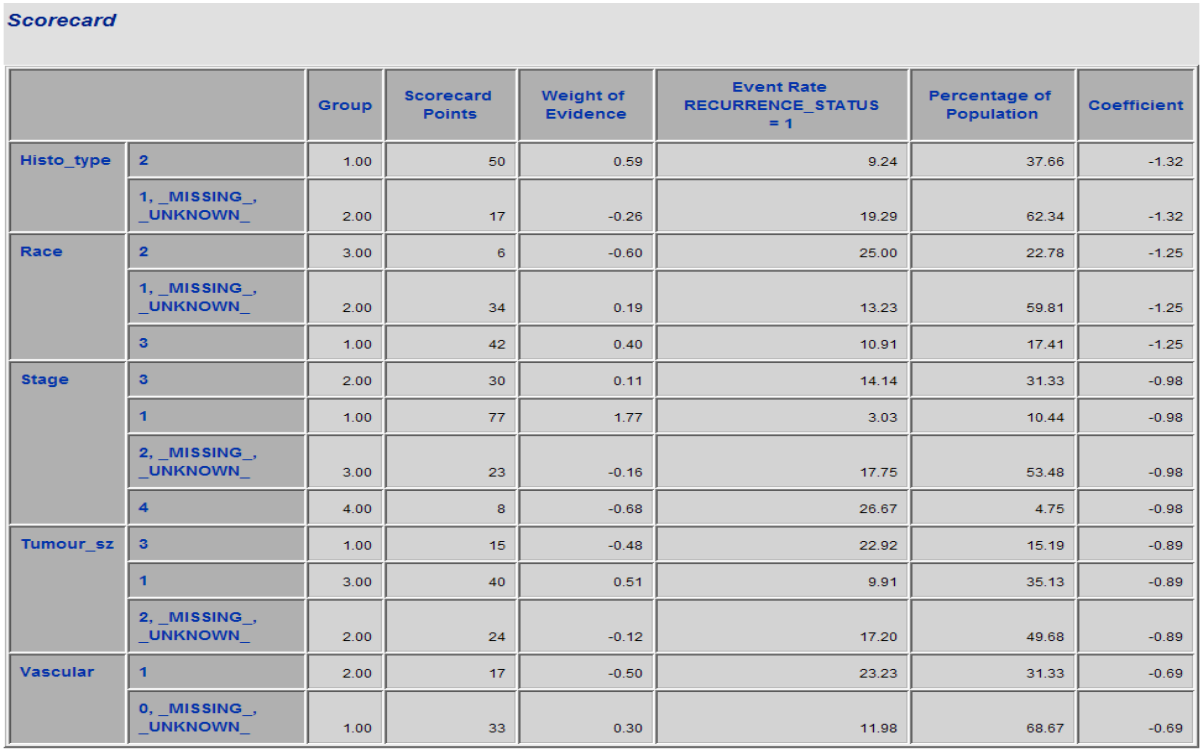
Display 5. hows the value of odds, scorecard points and points of double odds. The value of odds was 50, scorecard points (score) was 200 and points of double odds was 20.



Property	Value
General	
Node ID	Scorecard
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Scorecard Points	...
Score Ranges	...
Analysis Variables	WOE
Freeze Scorecard Poi	None
Publish Score Code	
Output Variables	Complete
Scaling Options	
Intercept Based Scor	No
Reverse Scorecard	No
Odds	50.0
Scorecard Points	200.0
Points to Double Odd	20.0
Scorecard Type	Detailed
Precision	0
Bucketing Method	Min/Max Distribution
Number of Buckets	25
Revenue Accepted G	1000
Cost Accepted Bad	50000
Current Approval Rat	70.0
Current Event Rate	2.5
Generate Characteris	No

Display 5. Screenshot of Scorecard Node Property

Display 6 shows the results of BCR-scorecard model that consist of scorecard point for each attribute and also the WOE values.



Scorecard

		Group	Scorecard Points	Weight of Evidence	Event Rate RECURRENCE_STATUS = 1	Percentage of Population	Coefficient
Histo_type	2	1.00	50	0.59	9.24	37.66	-1.32
	1, _MISSING_, _UNKNOWN_	2.00	17	-0.26	19.29	62.34	-1.32
Race	2	3.00	6	-0.60	25.00	22.78	-1.25
	1, _MISSING_, _UNKNOWN_	2.00	34	0.19	13.23	59.81	-1.25
	3	1.00	42	0.40	10.91	17.41	-1.25
Stage	3	2.00	30	0.11	14.14	31.33	-0.98
	1	1.00	77	1.77	3.03	10.44	-0.98
	2, _MISSING_, _UNKNOWN_	3.00	23	-0.16	17.75	53.48	-0.98
	4	4.00	8	-0.68	26.67	4.75	-0.98
Tumour_sz	3	1.00	15	-0.48	22.92	15.19	-0.89
	1	3.00	40	0.51	9.91	35.13	-0.89
	2, _MISSING_, _UNKNOWN_	2.00	24	-0.12	17.20	49.68	-0.89
Vascular	1	2.00	17	-0.50	23.23	31.33	-0.69
	0, _MISSING_, _UNKNOWN_	1.00	33	0.30	11.98	68.67	-0.69

Display 6. Screenshot of BCR-Scorecard Node

Table 4 summarizes the value of IV, WOE and Scorecard Points in this study. For the variables with IV greater than 0.10 are selected as input in building the BCR-scorecard model for the training data set. Thus, Stage, histological type, vascular invasion, race, and tumour size are included in the BCR-scorecard model.

Based on WOE, patients with low risk are those with attributes such as stage I (1.77031), stage III (0.10817), lobular infiltrating (0.58881), absent of vascular invasion (0.29875), Indian patients (0.40463), Malay patients (0.18556) and tumour size less than 3cm (0.51185). Meanwhile, patients with high risk are those with attributes such stage II (-0.16215), stage IV (-0.68383), ductal infiltrating (-0.26411), present of vascular invasion (-0.50019), Chinese patients (-0.59682) tumour size 6cm and above (-0.48241) and tumour size between 3cm and 5.9cm (-0.12373).

The lowest scorecard point for each attribute of every variable indicates higher risk of recurrence individually. Results show that the highest risk factors of recurrence were stage IV (8), ductal infiltrating carcinoma (17), present of vascular invasion (17), Chinese patients (6) and tumour size 6cm and above (15).

Variables	Groups	Attributes	IV	WOE	Scorecard Points
Stage	1	I	0.222	1.77031	77
	2	III		0.10817	30
	3	II		-0.16215	23
	4	IV		-0.68383	8
Histological type	1	Lobular infiltrating	0.154	0.58881	50
	2	Ductal infiltrating		-0.26411	17
Vascular invasion	1	Absent	0.148	0.29875	33
	2	Present		-0.50019	17
Race	1	Indian	0.143	0.40463	42
	2	Malay		0.18556	34
	3	Chinese		-0.59682	6
Tumour size	1	≥ 6cm	0.126	-0.48241	15
	2	3cm-5.9cm		-0.12373	24
	3	<3cm		0.51185	40

Table 4. BCR-Scorecard Model Results

Referring to Table 5, let us consider two patients. Patient A has the following attributes: Malay, stage IV, tumour size between 3cm and 5.9cm and ductal carcinoma. Patient B has the following attributes: Indian, lobular infiltrating, absent of vascular invasion and stage III. The calculation of total scores show that Patient B have low risk of recurrence since the score calculated is greater than 140 (155>140) and unfortunately patient A will have high risk of recurrence since the score obtained is less than 140 (92<140).

Patient	Calculation	Score
A	34 + 8+ 33+ 17	92
B	42 + 50 + 33+ 30	155

Table 5. Calculation of BCR Score Example

B. Logistic Regression Model Result

Table 6 is presents the Analysis of Maximum Likelihood Estimates results for the Logistic Regression Model. Results show that out of thirteen variables, there are five variables which were significant: age group, race, stage, histological type and progesterone receptors. The odds-ratios are shown in the last column. Odd-ratio which are greater than one indicates higher risk of recurrence.

The odds ratio for age group indicates younger patients (age <40 years old) are at higher risk compared to older patients (60 and above years old). Chinese patients are at higher risk compared to Indian. Patients diagnosed at stage II are at higher risk compared to stage IV. Results show that ductal infiltrating carcinoma patients have higher risk compared to lobular infiltrating carcinoma patients. Finally, patients with positive progesterone receptors are at higher risk compared to patients with negative progesterone receptors. The common factors for the BCR-scorecard and logistic regression are: stage, histological type and race.

Variable	Attributes	Wald Chi Square	p-value	Exp (Est)
Intercept		14.79	0.0001	0.142
Age group	<40 years	5.03	0.0249*	2.155
	40–59 years	2.13	0.1449	0.702
Marital status	Single	1.80	0.1793	1.382
Race	Malay	3.11	0.0776	0.638
	Chinese	7.98	0.0047**	2.276
Family history	Yes	1.69	0.1941	1.492
Stage	I	3.07	0.0799	0.232
	II	4.05	0.0442*	2.195
	III	0.00	0.9511	0.975

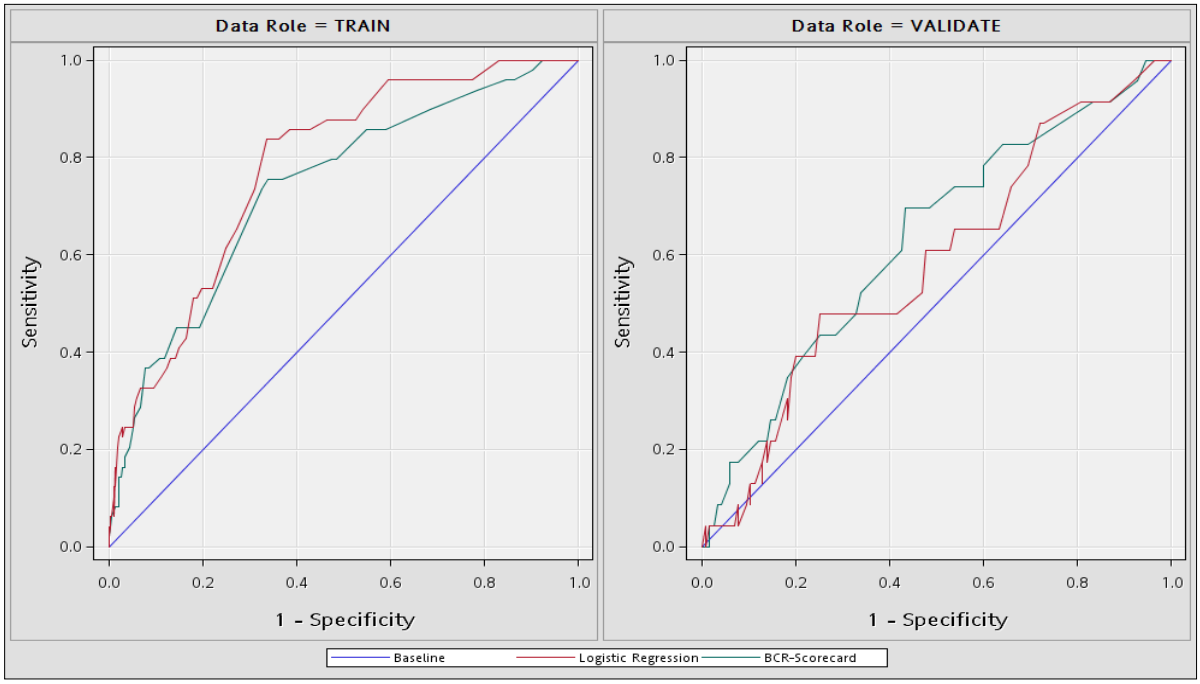
Tumour size	<3cm	1.36	0.2438	0.705
	3cm-5.9cm	0.25	0.6172	1.130
Histological grade	Grade I	0.13	0.7162	1.131
	Grade II	0.43	0.5117	0.849
Histological type	Ductal infiltrating	9.29	0.0023**	1.887
Lymph nodes status	0	0.34	0.5572	0.844
	1-3	0.17	0.6765	0.871
	4-9	1.37	0.2426	1.510
Vascular invasion	Present	3.30	0.0693	0.707
Estrogens receptors	Positive	1.48	0.2240	0.750
Progesterone receptors	Positive	4.66	0.0308*	1.665
CerB2	Positive	0.44	0.5078	1.127

Note. *p<0.05 **p<0

Table 6. Logistic Regression Results

C. Model Comparison Result

After building the BCR-scorecard and logistic regression model, classification performance were compared using Receiver Operating Characteristic (ROC) curve, misclassification rate, sensitivity and specificity. Display 7 displays the Receiver Operating Characteristic (ROC) Curve. By looking at ROC curve of validation sample, the ROC plot of green line (BCR-scorecard model) is higher than the red line (Logistic Regression model). This shows that BCR-scorecard model have higher sensitivity compared to the Logistic Regression model.



Display 7. Receiver Operating Characteristic (ROC) Curve

Based on Table 7, the BCR-scorecard has lower misclassification rate (18.12%) compared to logistic regression model (23.19%.) The accuracy rate for BCR-scorecard model is 82% and higher than for logistic regression model (77%). However, the sensitivity (13.04%) of the logistic regression model is higher than the scorecard model (0.04%).

	Models			
	BCR-scorecard		Logistic Regression	
	Train	Valid	Train	Valid
Misclassification	0.155063	**0.181159**	0.142405	0.231884
Accuracy	0.844937	**0.818841**	0.857595	0.768116
Sensitivity	0.081633	0.043478	0.163265	**0.130435**
Specificity	0.985019	**0.973913**	0.985019	0.895652

Table 7. Comparison Models Result

CONCLUSION

From this study, although the BCR-scorecard model has better predictive ability with lower misclassification rate (18%) compared to Logistic Regression model (23%), the sensitivity of the BCR-scorecard model is still low possibility due to the small sample size and small numbers of risk factors. Five important risk factors were identified: histological type, race, stage, tumour size and vascular invasion in predicting recurrence status. This study is the first attempt to use the concept of credit scoring for medical applications. The study has shown the feasibility of using a medical scorecard for classifying patients. The main limitation of this study is the occurrence of high percentage of missing values which led to the small sample size of study. With larger sample size and the involvement of breast cancer specialists, a more in-depth and comprehensive understanding of breast cancer recurrence can be obtained.

REFERENCES

[1] American Cancer Society. Cancer Facts & Figures. (2010). Atlanta: American Cancer Society. Available at: http://www.cancer.org

[2] American Cancer Society. Cancer Facts & Figures. (2012). Atlanta: American Cancer Society, Inc. Available at: http://www.cancer.org

[3] American Cancer Society. Breast Cancer Facts & Figures. (2011-2012). Atlanta: American Cancer Society, Inc. Available at: http://www.cancer.org

[4] Yip, C.H. (2012). Breast Cancer in Asia. Available at: www.biosensingtech.co.uk

[5] Omar, Z.A., and Tamin, N.S.I. (2011). National Cancer Registry Report Malaysia Cancer Statistics-Data and Figure 2007. National Cancer Registry, Ministry of Health Malaysia.

[6] Globocan. (2008). Cancer incidence, mortality and prevalence worldwide in 2008. Available at: http://globocan.iarc.fr

[7] Imaginis. (2000). The Breast Cancer Resource. Breast Cancer Recurrence (2000). Available at: http://www.imaginis.com/breasthealth/bcrecurrence.asp

[8] Preliminary Report. (2008). Clinical Research Centre. National Cancer Patient Registry (NCPR).

[9] Norsa'adah, B., Rusli, B.N., Imran, A.K., Naing, I., and Winn T. (2005). Risk factors of breast cancer in women in Kelantan, Malaysia. *Singapore Med J 2005; 46(12):698-705.*

[10] Mohammed, R.A., Isa, Z.M, Shah S.A, Mohd Nor, M.I., Chen, R, Ismail, F., and Radman, S.A. (2009). Eight year survival among breast cancer Malaysian women from University Kebangsaan Malaysia Medical Centre. *Asian Pacific J Cancer Prev, 10: 1075-1078.*

[11] Ong, T.A., and Yip, C.H. (2003). Short-term survival in breast cancer: The experience of the University of Malaya Medical Centre. *Asian J.Surg, 26, 169-75.*

[12] Breast Cancer Care. (2011). Breast Cancer in Families. Available at:www.breastcancercare.org.uk

[13] Yip, C.H., Mohd Taib, N.A. and Mohamed, I. (2006). Epidemiology of Breast Cancer in Malaysia. *Asian Pacific Journal of Cancer Prevention. 7, 369-374.*

[14] Hisham, A.N., and Yip, C.H. (2004). Overview of breast cancer in Malaysian women: A problem with late diagnosis. *Asian Journal of Surgery 27(2), pp. 130-133.*

[15] Hisham, A.N., and Yip, C.H. (2003). Spectrum of breast cancer in Malaysian women: Overview. *World J. Surg. 27(8).*

[16] Voogd, A.C., Nielsen, M., Peterse, J.L., Blichert-Toft, M., Bartelink, H., Overgaard, M., Van Tienhoven, G. (2001). Differences in risk factors for local and distant recurrence after breast-conserving therapy or mastectomy for stage I and II breast cancer: Pooled results of two large European randomized trials. *Journal of Clinical Oncology, 19(6), 1688-1697.*

[17] Breast Cancer Risk and Risk Factors. (2010). Available at: http://www.breastcancer.org

[18] Shadiya Mohamed, S. Baqutayan, Gogilawani, Akbariah Mohd Mahdzir, Saidatul Sariyah. (2012). Causes of breast cancer: comparison between the three races in Malaysia. *Journal of Health Sciences (J Health Sci) 2012; 2(2): 19–29.*

[19] Pathy, N.B., Yip, C.H., Taib, N.A., et al. (2011). Breast cancer in a multi ethnic Asian setting: Results from the Singapore-Malaysia hospital-based breast cancer Registry. *Breast, 20, 75-80.*

[21] Brewster, A.B., Do, K.A., Thompson, P.A., Hahn, K.M., Sahin, A.A., Cao, Y., et al. (2007). Relationship between epidemiologic risk factors and breast cancer recurrence. *J Clin Oncol 25:4438-4444. American Society of Clinical Oncology.*

[22] O'Leary, K., Estabrooks, C., Olson, K., and Cumming, C. (2007). Information acquisition for women facing surgical treatment for breast cancer: Influencing factors and selected outcomes. *Patient education and counselling, 69, 5-19.*

[23] Brewster, A.B., Hortobagyi, G.N., Broglio, K.R., Kau, S.W., et al. (2008). Residual risk of breast cancer recurrence 5 years after adjuvant therapy. *J Natl Cancer Inst 2008; 100: 1179-1183.*

[24] Society for Women Health Research. (2010). Life After Early Breast Cancer (ABC) Disease Awareness Initiative. Risk of recurrence in early breast cancer.
Available at: http://www.lifeabc.org/risk_recurrence_more.html

[25] Botteri, E., Bagnardi, V., Rotmensz, N., Gentilini, O., Disalvatore, D., et al. (2010). Analysis of local and regional recurrences in breast cancer after conservative surgery. *Annals of Oncology 21: 723–728.*

[26] Farhana Badar, Imran Moid, Farah Waheed, Anbreen Zaidi, Bilal Naqvi, and Syakeeb Yunus. (2005). Variables associated with recurrence in breast cancer patients-the Shaukat Khanum Memorial Experience. *Asian Pasific Journal of Cancer Prevention, Vol 6.*

[27] Chapman, J.W., Fish, E.B., and Link, M.A. (1999). Competing risks analyses for recurrence from primary breast cancer. *British Journal of Cancer (1999) 79(9/10), 1508–1513.*

[28] Jagsi, R., Phil, D., Raad, R.A., Goldberg, S., Sullivan, T., Michaelson, J., Powell, S.N., and Taghian, A.G. (2005). Locoregional recurrence rates and prognostic factors for failure in node-negative patients treated with mastectomy: implications for postmastectomy radiation. *Int. J. Radiation Oncology Biol. Phys., Vol. 62, No. 4, pp. 1035–1039.*

[29] Ozlem, E.R., Coskun, H.S.,, Soyuer I., Mustafa Altinbas. (2001). Late onset of distant metastases from breast cancer. *Turkish Journal of Cancer Vol. 31/ No. 3/ 2001.*
[30] Shigematsu, H., Kawaguchi, H., Nakamura, Y., Tanaka, K., Shiotani, S., et al. (2011). Significant survival improvement of patients with recurrent breast cancer in the periods 2001-2008 vs. 1992-2000. *BMC Cancer 2011, 11:118.*
[31] Martin, M., Gonzalez-Palacios, F., Cortes, J., De La Haba, J., and Schneider, J. (2009). Prognostic and predictive factors and genetic analysis of early breast cancer. *Clinical and Translational Oncology, 11(10), 634-42.*
[32] Mester, L. (1997). What's the Point of Credit Scoring? *Business Review.*
[33] Caire, D., and Kossmann, R. (2003). Credit Scoring: Is It Right For Your Bank?
[34] SAS institute Inc. (2009). SAS Notes. Building Credit Scorecards Using Credit Scoring for SAS Enterprise Miner. *A SAS Best Practice Paper.*
[35] Koh, H.C., Tan, W.C., and Goh, C.P. (2004). Credit scoring using data mining techniques. *Singapore Management Review. Vol.26, No.2, pp.25-47.*
[36] Yap, B. W., Ong, S. H, & Nor Huselina, M. H. (2011). Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models, *Expert Systems with Applications*, 38, 13274-13283

ACKNOWLEDGEMENT

The authors acknowledge Universiti Teknologi MARA for its support and facilities and we would like to express our gratitude to hospital Kuala Lumpur (HKL) for their contributions by providing the data. We also thank SAS Institute Inc. for the funding of this study to be presented in SAS Global Forum 2014.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Nurul Husna Jamian
Organization: Universiti Teknologi MARA (UiTM)
Address: Universiti Teknologi MARA,
40450 Shah Alam, Selangor,
Malaysia.
Phone: +060102359879
Email: nurul872@perak.uitm.edu.my

Name: Yap Bee Wah
Organization: Universiti Teknologi MARA (UiTM)
Address: Universiti Teknologi MARA,
40450 Shah Alam, Selangor,
Malaysia.
Phone: +060126344464
Email: beewah@tmsk.uitm.edu.my

Name: Nor Aina Emran
Organization: Hospital Kuala Lumpur (HKL)
Address: Department of Surgery,
Hospital Kuala Lumpur,
50586 Kuala Lumpur, Malaysia.
Phone: +060192750777
Email: aina_emran@yahoo.com