Saudi Journal of Biological Sciences 27 (2020) 1624–1631

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Prospective breast cancer risk factors prediction in Saudi women

Sawsan Babiker [a,b,*], Omaima Nasir [c], S.H. Alotaibi [d], Alaa Marzogi [e], Mohammad Bogari [e], Tahani Alghamdi [e]

[a] Department of Mathematics, Turabah University College, Taif University, Saudi Arabia
[b] Department of Mathematics, Faculty of Sciences, Gezira University, Sudan
[c] Department of Biology, Turabah University College, Taif University, Saudi Arabia
[d] Department of Chemistry, Turabah University College, Taif University, Saudi Arabia
[e] King Abd Alla Medical Centre, Mecca, Saudi Arabia

ARTICLE INFO

Article history:
Received 5 January 2020
Revised 9 February 2020
Accepted 23 February 2020
Available online 2 March 2020

Keywords:
Breast cancer
Risk Factors
Logistic Regression
Saudi women

ABSTRACT

Women's health is affected by breast cancer worldwide and Saudi Arabia (SA) is no exception. Malignancy has enormous consequences for social, psychological and public health. The aim of this study was to examine the risk factors for Saudi women from breast cancer using logistic regression models. In 135 patient cases for different stages of breast cancer was used to study case management, 270 healthy women from King Abd Alla Medical City, Mecca, SA were taken to predict the probability of women developing breast cancer, logistic regression was analyzed taking factors such as age, marital status, family history, parity, age at first full-term pregnancy, menopausal status, body mass index (BMI) and breast feeding. The logistic regression model showed that there are important risk factors (age, marital status, family history, parity, age at first full-term pregnancy, menopausal status, body mass index, and breast feeding) in development of breast cancer. Fewer cases were observed in unmarried women, age $\leq$30, BMI $\leq$20. In contrast, more cases were found with women age 41–50 married, BMI > 30, a smaller number of children, not breast feeding, age of first pregnancy $\geq$30, menopausal status age at 46–50. Based on our data there is role of risk factors in developing breast cancer, less BMI, the increase number of children, breast feeding, which are playing as protective factor for breast cancer.

1. Introduction

The second most common cancer in the world is Breast Cancer (BC). United States Cancer Society report showed that about 1.3 million American women were diagnosed with BC and 0,5 million death each year because of malignancies (American Cancer Society, 2012). The increased number of BC cases reported from different hospitals in most previous epidemic studies observed in Saudi Arabia (SA), although female breast cancer in SA has a lower incidence. Saudi Cancer Registry has confirmed that female breast cancer was the most prevalent Saudi cancer in the 14 years period (1994 – 2008) compared to other developed countries (Saudi Cancer Registry, 2005, 2008, Al-Qahtani 2007; Ravichandran et al., 2009). The awareness of the risk factor of breast cancer is also slightly adequate, which could have a significant impact on many etiological factors, including genetic, reproductive, ecological and socioeconomic factors. In Arab countries most of these variables were not explored in depth (Salah et al., 2010; Al Diab et al., 2013), we have sought to identify the highest risk factors for breast cancer with standard logistic regression models, which have been used for data analysis in many areas over the past decade.

Here, logistical concepts are briefly described as statistics with a fast-logistical distribution function account, simple logistic regression analysis, multiple logistics regression models, coefficients meaning testing and confidence interval assessment. Furthermore we define how the optimal logistic regression model is to choose variables that result in a "best" model in the empirical contexts of the problem, and how best to interpret the data and match the estimated logistic regression model, In our study population, we have predicted the most significant BC risk factors which could help to develop BC risk reduction strategies.

* Corresponding author at: Department of Mathematics , Turabah University College , Taif University, 21995, Saudi Arabia.
E-mail address: sawsanbabiker@gmail.com (S. Babiker).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.sjbs.2020.02.012
1319-562X/

2. Materials and methods

2.1. Study design

From January 2017 to December 2017, the case management study was conducted in King Abd Alla Medical Centre (KAMC), Mecca, SA. The incident case of patient admitted in the KAMC due to diagnosis with breast cancer were chosen for the study, all women confirmed diagnosis with breast cancer were interviewed by one investigator. For access to the corresponding KAMC information, written consent was obtained from the Supervisor of KAMC Review Board for all cases and control samples included in the analysis and no direct contact was established.

2.2. Case sample

Breast cancer patient's records at KAMC, from January 2017 up to December 2017, were chosen. Data collected throw questionnaire including socio-demographic factors (age, and marital status), reproductive factors (parity, age at first pregnancy, menopausal status, and breast-feeding) and body mass index (BMI). In addition to specialist and pathology records from which risk factors can be identified, the data collection of patients with breast cancer is accomplished by analyzing patient information through a direct interview between the patient and the related clinician.

2.3. Control sample

The control women were recruited randomly, residing in the same geographical region and admitted to the KAMC without a history of breast problems or neoplastic diseases. The demographic and risk factors data were collected by means of interview schedule, including information about the control same as in cases.

2.4. Data set

Following approval from the reviewing committee, the data for this analysis were obtained from KAMC. The National Institutes of Health accredited all researchers to protect participants in human research.

This study was conducted based on a sample of 405 people, including 135 cases (patients with breast cancer) and 270 control cases (not patients with breast cancer). Among women with breast cancer, 112 (83.0%) and 217 (80.4%) control are married. There were socio-demographic (age, and marital) factors, reproductive (parity, first pregnancy age, menopausal status, breast-feeding) and BMI as the risk factors assessed for the model's adaptation.

2.5. Methods

We have followed Salah et al. (2010) methods. The relationship between a binary variable and one or more explanatory values is defined by the logistic regression method (Appendix A) according to Cox and Snell (1989), Concato et al. (1993), Collaborative Group, (2001), Ravichandran (2005), Salah et al. (2010), and Al Diab et al. (2013).

2.6. Statistical analysis

Logistic regression helps to model the probability of women developing BC based on social-demographic (age and marital status), reproductive (parity, age at first birth, menopausal status and breast-feeding) and BMI variables. These variables are calculated according to Table 1. The research was conducted on the predictive effect of each variable in relation to breast cancer risk in order to calculate odds ratio (OR) and 95% confidence intervals (CI), as illustrated by Tables 6–9 of the (Appendix A), Eqs. (1)–(4) of (Appendix B) (Austin and Tu, 2004; Hadjisavvas et al., 2010), and Eq. (5) of (Appendix C), (Collett, 1991; Hosmer et al., 2000; Bagley Steven et al., 2001, Austin and Tu, 2004; Hadjisavvas et al., 2010; Genuer et al., 2010, Yusuff et al., 2012; Elkum et al., 2014). Risk factors associated with breast cancer have been entered into a multivariate logistic regression analysis of the forward-looking range.

3. Results

3.1. Socio-demographic factors

3.1.1. Age

Breast cancer cases and controls were detected in patients as young as 29 years and as old as 69 years with a mean ± S.E. 46.5 ± 0.573 and 45.5 ± 0.433 years for cases and controls, respectively as shown in (Table 1).

3.1.2. Marital status

Distribution of patients and controls according to the age group and marital status shown in Table 1, and Fig. 1. 5.2% comprised of breast cancer cases and only 5.6% of control subjects respondents were married (Table 1).

Results from (Table 1), shows that the maximum risk factors is in the age group of 41 to 51 with the cases of 65 out of 129 control samples, followed by 55 cases of breast cancer from the age group

Table 1
Frequency distribution of socio-demographic factors.

Variables	Case No (%)	Control No (%)	χ^2	P -value
Age group (years old)			18.968	0.000
30 and less	2 (1.5)	4 (1.5)		
31– 40	9 (6.7)	21 (7.8)		
41–50	65 (48.1)	129 (47.8)		
51– 60	55 (40.7)	109 (40.4)		
Above 60	4 (3.0)	7 (2.6)		
Marital status			13.452	0.001
Single	7 (5.2)	15 (5.6)		
Married	112 (83.0)	217 (80.4)		
Divorced	6 (4.4)	14 (5.2)		
Widowed	10 (7.4)	24 (8.8)		

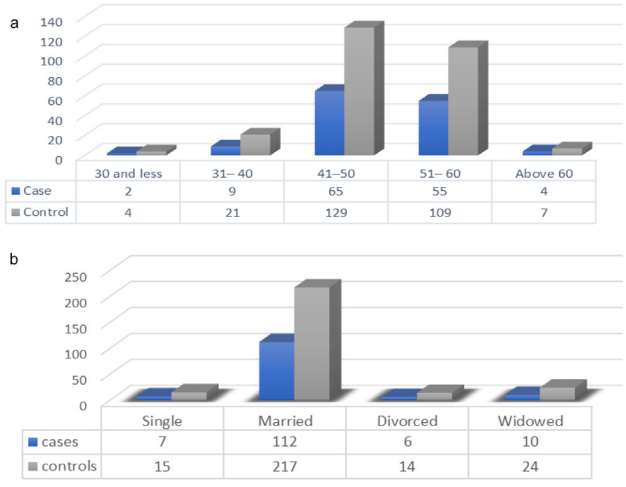

Fig. 1. Distribution of patients and controls according to age group (1.a) and marital status (1.b).

of 51 to 60 out of 109 controls and less case of 2 out of 4 was observed in the age group with less than 30. However, results of risk factor such as marital status was observed in married cases were high with 112 cases out of 217, compared to 6 cases with divorced out of 14, single risk factor had least 7 cases out of 15 control and 10 widowed cases out of 24 control.

As shown in (Table 2), BMI there had a significance p-value (0.000) in which (42.3%) of cases were obese, whereas 30.8% of control subjects were obese (Fig. 2).

The more cases were observed with the BMI > 30 with the cases of 57 out of 83, 50 cases out of 63 controls were observed with the BMI 25–29, followed by the cases with 27 out of 75 with BMI 20–24, only one case was observed out of 40 controls with the BMI < 20.

Significance p-value was shown in Table 3 regarding the distribution of patients and the controls group. Our results suggest that women with more number of children like > 10 was observed with 2 cases taking from 78 controls, 54 cases was observed from 49 with women bearing 5–10 children, 42 cases of breast cancer from 50 women having 1–4 children, 37 cases out of 93 controls were observed with women with 0 number of children Fig. 3.

Distribution of patients and controls according to breast-feeding as shown in (Table 4), and (Fig. 4). According to the case study the women doing breast feeding were observed with cases No.53 and 106 were observed as controls, whereas 82 women cases was observed out of 164 controls with no breast feeding.

In Table 5, and Fig. 5, show the distribution of patients and controls according to reproductive variables as first pregnancy, family history, and menopausal state. For age ≤30, the 84 cases were observed with 148 control, in the variables at age ≥30 only 14 cases were observed with 29 control, while in nulliparous 37 cases out of 93 were observed with no significant difference (P-value >0.05). For the family with breast cancer history, 36 cases out of 20, whereas females with no family history showed 48 cases out of 111 with high significant difference (P-value <0.05). As well as, the risk factor such as menopausal no significant difference (P-value >0.05) as for age ≤ 45 women were 5 out of 8, women with 46–50 years cases were 51 out of 104, and women > 50 years showed 28 cases out of 67 control.

All variables show significance variation, (Table 6), by using Model -1 as follows:

$$Logit\left(\hat{P}\right) = -3.173 + 1.488\text{Age} + 0.725\text{FH} - 1.20\text{MnS} - 0.998\text{MS} - 0.697\text{P} + 0.662\text{AP} + 0.416\text{BMI} \quad (1)$$

By using fitted Model-2, the age at first pregnancy shows significance variation with P-value (0.014) other variables were not significance (Table 7).

$$Logit\left(\hat{P}\right) = -0.069 - 5.458\ \text{AP} + 1.381\text{P} - 1.496BF \quad (2)$$

Table 2
Distribution of patients and controls according to body mass index (BMI).

Variables	Case No (%)	Control No (%)	χ^2	P -value
BMI at diagnoses			33.74	0.000
<20	1 (0.7)	49 (18.1)		
20–24	27 (20.0)	75 (27.8)		
25–29	50 (37.0)	63 (23.3)		
>30	57 (42.3)	83 (30.8)		

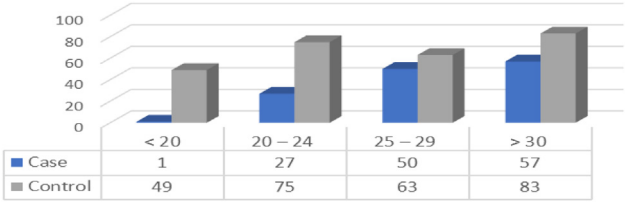

Fig. 2. Distribution of patients and controls according to BMI.

Table 3
Distribution of patients and controls according to Number of children.

Variables	Case No (%)	Control No (%)	χ^2	P -value
Number of children				
0	37 (27.4)	93 (34.4)	30.691	0.000
1–4	42 (31.1)	50 (18.5)		
5–10	54 (40.0)	49 (18.1)		
>10	2 (1.5)	78 (28.9)		

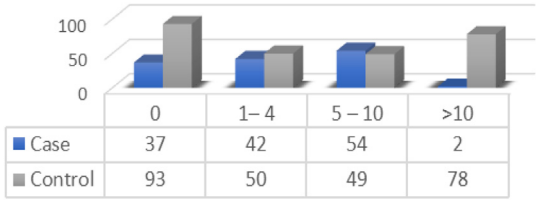

Fig. 3. Distribution of patients and controls according to number of children.

Table 4
Distribution of patients and controls according to breast- feeding.

Breast feeding	Case N (%)	Control N (%)	χ^2	P -value
No	82 (60.7)	164 (60.7)	0.727	0.394
Yes	53 (39.3)	106 (9.3)		

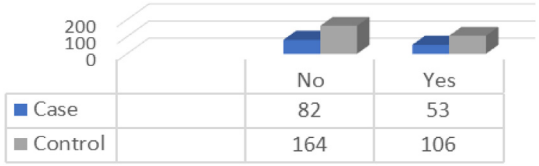

Fig. 4. Distribution of patients and controls according to breast-feeding.

Table 5
Distribution of patients and controls according to reproductive variables.

Variables	Case N (%)	Control N (%)	χ^2	P -value
Age at first pregnancy				
≤30	84 (62.2)	148 (54.8)	2.237	0.271
>30	14 (10.4)	29 (10.7)		
Nulliparous	37 (27.4)	93 (34.4)		
Family history				
Yes	36 (26.7)	20 (7.4)	31.101	0.000
No	48 (35.5)	111 (41.1)		
Not sure	51 (37.8)	139 (51.5)		
Menopausal status				
≤45	5 (3.7)	8 (3.0)	4.364	0.060
46–50 years	51 (37.8)	104 (38.5)		
>50 years	28 (20.7)	67 (24.8)		
Not sure	51 (37.8)	91 (33.7)		

Table 6
Variable in Model -1.

Variable	$\hat{\beta}$	SE ($\hat{\beta}$)	Wald	P-value	$\hat{OR}$	95% CI$\hat{OR}$	
						Lower	Upper
Age group	1.448	0.357	16.493	0.000	4.254	2.115	8.555
Marital status (MS)	−0.998	0.468	4.543	0.033	0.369	0.147	0.923
BMI	0.416	0.172	5.870	0.015	1.515	1.083	2.121
Family history (FH)	0.725	0.178	16.499	0.000	2.064	1.455	2.927
Age of first pregnancy (AP)	0.662	0.346	3.649	0.056	1.938	0.983	3.823
Parity (P)	−0.697	0.232	9.077	0.003	0.498	0.316	0.784
Menopausal status (MnS)	−1.120	0.355	9.960	0.002	0.326	0.163	0.654
Constant	−3.173	1.346	5.559	0.018	0.042		

Table 7
Variables in Model -2.

Variable	$\hat{\beta}$	S.E. ($\hat{\beta}$)	Wald	P-value	$\hat{OR}$	95% CI$\hat{OR}$	
						Lower	Upper
Age at first pregnancy (AP)	−5.458	2.221	6.041	0.014	0.004	0.000	0.331
Parity (P)	1.381	0.813	2.885	0.089	3.978	0.809	19.569
Breast feeding (BF)	−1.496	0.903	2.749	0.097	0.224	0.038	1.313
Constant	−0.069	1.202	0.003	0.955	0.934		

Table 8
Variables in Model-3.

Variable	$\hat{\beta}$	S.E. ($\hat{\beta}$)	Wald	P-value	$\hat{OR}$	95% CI$\hat{OR}$	
						Lower	Upper
Menopausal status	0.599	0.146	16.885	0.000	1.821	1.368	2.424
Constant	−0.795	0.166	22.917	0.000	0.452		

Table 9
Model assessment.

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	213.013[a]	0.293	0.398
2	213.458[a]	0.292	0.396

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.906	8	0.0013
2	13.181	8	0.0014
3	1.998	6	0.00

[a]Estimation terminated at iteration number 6 because parameter estimates changed by<0.001.
[b]Estimation terminated at iteration number 6 because parameter estimates changed by<0.001.

By using fitted Model-3, the Menopausal status shows significance variation (Table 8). Logit $\left(\hat{P}\right) = -0.795 + 0.599MnS$ (2)

The evaluation of Model in (Table 9), showed that R^2 = 0.398, in addition, R^2 value was low and small, but showed statistically significant forecasts (P-value < 0.05). Important assumptions were made about the relationship between changes in predictor values and changes in response value. Regardless of the R^2, the mean change in the answer for a unit of predictor change always reflects the relevant coefficients while other predictors are constant in the model. This type of information will certainly be of enormous value.

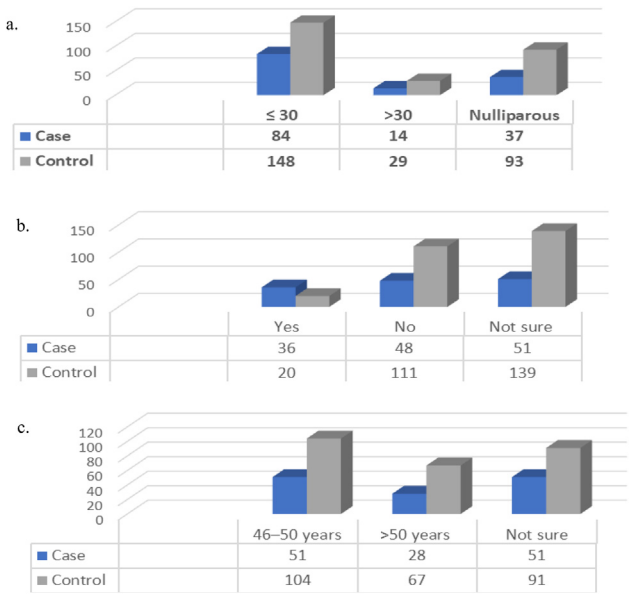

Fig. 5. Distribution of patients and controls according to age at first pregnancy (5.a), family history (5.b) and menopausal status (5.c).

4. Discussion

Backward elimination was conducted using SPSS version 21 software (SPSS, Inc., Chicago, IL, USA), and logistic regression was analyzed to the factors such as socio-demographic (age and marital status), reproductive (parity, age at first pregnancy, menopausal status and breast-feeding), and BMI. By using logistic regression models, we have found that there is a significant correlation between the BMI and an increase in the number of cases of breast cancer (Hopper John et al., 2018), which means that obese women

can be at high risk for breast cancer and the results are an alignment with what has been stated by Elkum et al. (2014). In addition, mothers with more children played a protective role in our data on breast cancer. Family history, on the other hand, plays a significant role, as in most other reports (Collaborative Group, 2001; Elkum et al., 2014). Family history is a risk factor in previous studies (Braithwaite et al., 2018), and logistic regression model is one of the best models used to determine risk factors (Dawood Shaheenah et al., 2014).

In the current study, breast feeding did not play a protective role in breast cancer, since a smaller number of breast feeding cases were observed. Some studies suggest it is possible to prevent breast cancer by breast-feeding and some studies have shown that breast cancer risk does not affect lactation (Lipworth et al., 2000). Nevertheless, epidemiological studies have indicated that populations with normal long lactation periods pose low breast cancer risks (Lipworth et al., 2000). These conflicting results suggest that the effects of breast cancer risk factors are likely to be small. It is definitely of interest to consider how lactation could help to prevent breast cancer, as it is a modifiable risk factor. Understanding the role of lactation may help us to understand the etiology of a disease of immense importance for public health. The women bearing a greater number of children earlier reported in lowering the breast cancer (Dall and Biritt, 2017), also menopausal stages effect risk of breast cancer (Chang- Claude et al., 2007).

5. Conclusions

Based on our data and tables suggested that the risk factor for developing breast cancer was at age group of 41–50, those are married having BMI > 30, bearing less children, not breast feeding, having pregnancy at the age of ≥30, though showing family history and menopausal status at the age of 46–50 had more number of breast cancer cases, whereas women who are single age less than 30, BMI <20 has less cases of breast cancer, data also suggest us that the women bearing children >10 and also breast feeding plays as protective role in developing breast cancer, and also less number of cases were observed with menopausal status at the age ≤45.

Acknowledgements

This study was supported by grants from the Taif Scientific Research Centre, Taif University, SA [No. 1-437-5326]. The authors acknowledge the assistance of Mr. Sultan Almugbel, Mrs. Doha Morsi radiology technologist from King Abd Alla Medical Center, Mecca, SA for data collection.

Declaration of Competing Interest

Author(s) declare that all works are original and this manuscript has not been published in any other journals.

Limitations

We don't interview the subjects face to face, all the information retrieved from the patient's hospital records, their validity and standard are open for bias. Recall bias was also expected as regards to their date e.g. age, age of 1st pregnancy, number of children, breast feeding.

Appendix A. Methods

We followed the methods of Salah et al. (2010). The relation between the binary variable with one or more explicatory variables is defined by the logistic regression model. The purpose of research with logistic regression is the same as that with a linear regression model in which it is believed that the dependent variable is continuous or distinct. The response variable is usually dichotomous in logistic regression, where the response variable may take value 1 with success probability p or value 0. with probability of failure 1-p. This type of variable is known as a binary. The relationship between predictor and response variables in logistic regression is not a linear function; instead, a logistic regression function is used, given as (Austin and Tu, 2004; Hadjisavvas et al., 2010; Elkum et al., 2014).

$$P(x) = \frac{\exp(\beta_0 + \beta_i x)}{1 + \exp(\beta_0 + \beta_i x)} \tag{1}$$

The logit transformation is a transformation of P(x) which is central to our study of logistic regression. This transformation is defined, in terms of P(x), as follows:

$$g(x) = \mathrm{logit}(p(x)) = \ln \frac{p(x_i)}{1 - p(x_i)} = \beta_0 + \beta_i x \tag{2}$$

where β_o and β_i are the logistic intercept and coefficients, respectively.

The parameters in this model, $\beta_0 + \beta_i x$, can no longer be estimated by least squares, but are found using the maximum likelihood method. The probability of success vs. failure is determined by logistic regression; therefore, the results of the analysis are in the form of an odds ratio. Logistic regression also shows connections between variables and strengths. The Wald statistics are typically used to determine the value for each independent variable of the single logistic regression coefficients. The Wald statistic for the β_i coefficient is:

$$Wald = \left[\frac{\beta_i}{S.E.(\beta_i)}\right]^2 \tag{3}$$

This value is distributed as chi-square with 1 degree of freedom. The Wald statistic is the square of the (asymptotic) t-statistic. The Wald statistic can be used to calculate a confidence interval for βi. We can assert with 100 $(1 - \alpha)$% confidence that the true parameter lies in the interval with boundaries $\hat{\beta} \pm Z_{\alpha/2}(ASE)$, where ASE is the asymptotic standard error of logistic $\hat{\beta}$. Estimates of parameters are derived using the maximum likelihood principle; Hypothesis tests are therefore based on comparisons between probabilities or deviances of nested models. The probability ratio check uses the ratio of the maximized probability value for the complete model (L1) to the maximized probability function value for the simplified model (L0). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1) \tag{4}$$

This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistics for a model with a rear removal process. The reverse removal process seems to be the preferred method of exploratory tests where the study starts with an entire or saturated model and variables in an iterative process are removed from the model. After removing each variable, the model fit is tested to ensure it fits the data properly. If the model cannot remove any more variables, the analysis is complete (Yusuff et al., 2012).

Appendix B. Validation

The validation test was carried out to determine if the study of logistic regression was satisfactory. The estimated accurate case percentage from major samples must be equal to or greater than the actual sample percentage. For calculating the percentage of correct instances, validation uses the other sample data with the

same coefficient values as main data. First, the data were divided into two groups. In order to determine coefficient values, 80 percent of the first data group was used as the key data. For validating the main results, the second group comprised 20 percent of the samples. The probability of each example from the validated data was determined after the coefficient values were obtained from the main data. Probability was defined as:

$$P(Y=m)=\frac{\exp(g(x))}{1+\exp(g(x))} \tag{5}$$

The reference probability was defined as:

$$P(Y=0)=\frac{1}{1+\exp(g(x))},\ \text{with}\ g(x)=\beta_0+\sum_{i=1}^{n}\beta_i x_i$$

β_0 is the intercept coefficient values, whereas β_i is the coefficient value of each factor contributing to occurrence With the observed probability, the probability of each test has been cross-validated. The percentage of correct classification cases has been obtained for cross-validation. Next, The correct classification case percentages of validated data is equivalent to the correct classification case percentage of principal data. There were two groups of results. In deciding the logistic regression model, the first 110 samples were taken. To validate the pattern, the remaining samples were used. To assess the percentage of correct classification events, the verified findings were used (Concato et al., 1993).

Appendix C. Variable selection

It is critical that the model contains all relevant variables and does not start with more than the number of observations justified (Bagley Steven et al., 2001; Austin and Tu, 2004; Hadjisavvas et al., 2010). Additional variables typically produce a better model that fits the data for a dataset. Excessive variables, however, influence the model coefficient and help overfit the model. A complex model with many small variables will lead to less predictive power and make interpreting the results difficult. The statistical variable selection process is based on two procedures. Next, interactions are shown as product terms in the interaction study, which is a concept of the regression model and not a single predictor variable, but rather the product of two predictors (Hosmer et al., 2000; Genuer et al., 2010). Interaction experiments were carried out to determine each variable's important values. Co-linearity analysis is the second method. With the consequent lack of statistical significance the disparity associated with these coefficients increases (Collaborative Group, 2001). The study of co-linearity was based on essential interaction test values. Each variable must have significant values less than 0.20 (Hosmer et al., 2000), used in the study of the logistic regression model (Concato et al., 1993).

Appendix D

Distribution of Age groups (cases and control) according risk factors.

(see Tables 1a–1g).

Table 1a
Distribution of age groups according to age at 1st pregnancy.

Age at 1st pregnancy							
			Ever N (%)	**< 30 N (%)**	**>30 N (%)**	**Total N (%)**	P-value
Control	**Age groups**	**30 and less**	2 (50.0%)	1 (25%)	1 (25%)	4(100%)	
		31–40	6 (28.6%)	15 (71.4%)	0 (0%)	21(100%)	
		41–50	36 (27.9%)	87 (60.5%)	15 (11.6%)	129 (100%)	0.148
		51–60	48 (44.0%)	48 (44.0%)	13 (11.9%)	109(100%)	
		above 60	2 (28.6%)	5 (71.4%)	0 (0%)	7(100%)	
	Total		**94 (34.81%)**	**147 (54.4%)**	**29 (10.7%)**	**270(100%)**	
case	**Age group**	**30 and less**	2 (100.0%)	0 (0%)	0 (0%)	2 (100.0%)	
		31–40	3 (33.3%)	4 (44.4%)	2 (22.2%)	9 (100%)	
		41–50	16 (24.6%)	45 (69.2%)	4 (6.2%)	65 (100%)	
		51–60	13 (23.6%)	33 (60%)	9 (16.4%)	55 (100	0.195
		above 60	2 (50%)	2 (50%)	0 (0%)	4 (100%)	
	Total		**36 (26.7%)**	**84 (62.2%)**	**15 (11.1%)**	**135 (100%)**	

Table 1b
Distribution of Age groups according to BMI.

			BMI				
			< 20 N (%)	**20–24 N (%)**	**25–29 N (%)**	**>=30 N (%)**	**Total N (%)**	P -value
Control	**Age group**	**30 and less**	0 (0%)	1 (25%)	0 (0%)	3 (75%)	4(100%)	
		31–40	3 (14.3%)	4 (19%)	6 (28.6%)	8 (38.1%)	21(100%)	
		41–50	22 (17.1%)	37 (28.7%)	36 (27.9%)	34 (26.4%)	129 (100%)	0.119
		51–60	24 (22%)	32 (29.4%)	17 (15.6%)	36 (33%)	109(100%)	
		above 60	0 (0%)	2 (28.6%)	4 (57.1%)	1 (14.3%)	7(100%)	
	Total		**49 (18.1%)**	**76 (28.1%)**	**63 (23.3%)**	**82 (30.4%)**	**270(100%)**	
case	**Age groups**	**30 and less**	0 (0%)	1 (50%)	1 (50%)	0 (0%)	2 (100.0%)	
		31–40	0 (0%)	0 (0%)	4 (44.4%)	5 (55.6%)	9 (100%)	
		41–50	0 (0%)	14 (21.5%)	24 (36.9%)	27 (41.5%)	65 (100%)	
		51–60	1 (1.8%)	11 (20%)	21 (38.2%)	22 (40%)	55 (1 0 0)	0.509
		above 60	0 (0%)	1 (25%)	0 (0%)	3 (75%)	4 (100%)	
	Total		**1 (0.7%)**	**27 (20%)**	**50 (37%)**	**57 (42.2%)**	**135 (100%)**	

Table 1c
Distribution of age groups according to menopausal status.

			Menopausal Status				
			45 and less N (%)	**46–50 N (%)**	**Above 50 N (%)**	**Total N (%)**	P -value
Control	**Age groups**	**30 and less**	4 (100.0%)	0 (0.0%)	0 (0.0%)	4(100%)	
		31–40	16 (76.2%)	2 (9.5%)	3 14.3%)	21(100%)	
		41–50	58 (45.0%)	59 (45.7%)	12 (9.3%)	129(100%)	0.000
		51–60	60 (55.0%)	4 (3.7%)	45 (41.3%)	109(100%)	
		above 60	0 (0.0%)	1(14.3%)	6 (85.7%)	7(100%)	
	Total		**138 (51.1%)**	**66 (24.4%)**	**66 (24.4%)**	**270(100%)**	
case	**Age groups**	**30 and less**	2 (100.0%)	0 (0.0%)	0 (0.0%)	2 (100.0%)	
		31–40	9 (100%)	0 (0%)	0 (0%)	9 (100%)	
		41–50	32 (49.2%)	33 (50.8%)	0 (0.0%)	65 (100%)	
		51–60	0 (0%)	0 (0%)	55 (100%)	55 (100	0.000
		above 60	0 (0%)	0 (0%)	4 (100%)	4 (100%)	
	Total		**43 (31.9%)**	**33 (24.4%)**	**59 (43.7%)**	**135 (100%)**	

Table 1d
Distribution of age groups according to marital status.

			Marital Status					
			Married N (%)	**Single N (%)**	**Widow N (%)**	**Divorce N (%)**	**Total N (%)**	P -value
Control	**Age groups**	**30 and less**	3 (75%)	0 (0%)	0 (0%)	1 (25%)	4(100%)	
		31–40	15 (71.4%)	2 (9.5%)	0 (0%)	4 (19%)	21(100%)	
		41–50	111 (86%)	4 (3.1%)	7 (5.4%)	7 (5.4%)	129 (100%)	0.376
		51–60	81 (74.3%)	6 (5.5%)	5 (4.6%)	17 (15.6%)	109(100%)	
		above 60	7 (100%)	0 (0%)	0 (0%)	0 (0%)	7(100%)	
	Total		**217 (80.4%)**	**12 (4.4%)**	**12 (4.4%)**	**29 (10.7%)**	**270(100%)**	
case	**Age groups**	**30 and less**	0 (0%)	2 (100%)	0 (0%)	0 (0%)	2 (100.0%)	
		31–40	8 (88.9%)	1 (11.1%)	0 (0%)	0 (0%)	9 (100%)	
		41–50	61 (93.8%)	1 (1.5%)	3 (4.6%)	0 (0%)	65 (100%)	
		51–60	54 (98.2%)	0 (0%)	1 (1.8%)	0 (0%)	55 (1 0 0)	0.142
		above 60	3 (75%)	0 (0%)	1 (25%)	0 (0%)	4 (100%)	
	Total		**126 (93.3%)**	**4 (3%)**	**5 (3.7%)**	**0 (0%)**	**135 (100%)**	

Table 1e
Distribution of age groups according to breast feeding.

			Breast feeding			
			Yes N (%)	**NO** N (%)	**Total** N (%)	P -value
Control	**Age groups**	**30 and less**	2 (50%)	2 (50%)	4(100%)	
		31–40	13 (61.9%)	8 (38.1%)	21(100%)	
		41–50	80 (62%)	49 (38%)	129 (100%)	0.397
		51–60	66 (60.6%)	43 (39.4%)	109(100%)	
		above 60	4 (42.9%)	4 (57.1%)	7(100%)	
	Total		**164 (60.7%)**	**106 (39.3%)**	**270(100%)**	
case	**Age groups**	**30 and less**	1 (50%)	1 (50%)	2 (100.0%)	
		31–40	4 (44.4%)	5 (55.6%)	9 (100%)	
		41–50	43 (66.2%)	22 (33.8%)	65 (100%)	
		51–60	32 (58.2%)	23 (41.8%)	55 (1 0 0)	0.508
		above 60	2 (50%)	2 (50%)	4 (100%)	
	Total		**82 (60.7%)**	**53 (39.3%)**	**135 (100%)**	

Table 1f
Distribution of age groups according to number of children.

			No of children					
			Null N (%)	1–4 N (%)	5–10 N (%)	>10 N (%)	Total N (%)	P -value
Control	**Age groups**	**30 and less**	2 (50%)	0 (0%)	2 (50%)	0 (0%)	4(100%)	
		31–40	6 (28.6%)	2 (9.5%)	3 (14.3%)	10 (47.6%)	21(100%)	
		41–50	35 (27.1%)	28 (21.7%)	23 (17.8%)	43 (33.3%)	129 (100%)	0.014
		51–60	48 (44%)	19 (17.4%)	18 (16.5%)	24 (22%)	109(100%)	
		above 60	2 (28.6%)	1 (14.3%)	3 (42.9%)	1 (14.3%)	7(100%)	
	Total		**93 (34.4%)**	**50 (18.5%)**	**49 (18.1%)**	**78 (28.9%)**	**270(100%)**	
Case	**Age groups**	**30 and less**	2 (100%)	0 (0%)	0 (0%)	0 (0%)	2 (100.0%)	
		31–40	3 (33.3%)	6 (66.7%)	0 (0%)	0 (0%)	9 (100%)	
		41–50	17 (26.2%)	16 (24.6%)	31 (47.7%)	1 (1.5%)	65 (100%)	
		51–60	13 (23.6%)	19 (34.5%)	22 (40%)	1 (1.8%)	55 (1 0 0)	0.139
		above 60	2 (50%)	1 (25%)	1 (25%)	0 (0%)	4 (100%)	
	Total		**37 (27.4%)**	**42 (31.1%)**	**54 (40%)**	**2 (1.5%)**	**135 (100%)**	

Table 1g
Distribution of age groups according to family history of BC.

			Family history of BC			
			Yes N (%)	NO N (%)	Total N (%)	P -value
Control	**Age groups**	**30 and less**	4 (100%)	0 (0%)	4(100%)	
		31–40	19 (90.5%)	2 (9.5%)	21(100%)	
		41–50	110 (85.3%)	19 (14.7%)	129 (100%)	0.326
		51–60	93 (85.3%)	16 (14.7%)	109(100%)	
		above 60	7 (100%)	0 (0%)	7(100%)	
	Total		**233 (86.3%)**	**37 (13.7%)**	**270(100%)**	
Case	**Age groups**	**30 and less**	2 (100%)	0 (0%)	2 (100.0%)	
		31–40	5 (55.6%)	4 (44.4%)	9 (100%)	
		41–50	40 (61.5%)	25 (38.5%)	65 (100%)	
		51–60	32 (58.2%)	23 (41.8%	55 (1 0 0)	0.283
		above 60	3 (75%)	1 (25%)	4 (100%)	
	Total		**82 (60.7%)**	**53 (39.3%)**	**135 (100%)**	

References

Abdurrahman, Al Diab, Shoeb, Qureshi, Al Saleh Khalid, A., Al Qahtani Farjah, H., Aamer, Aleem, Alghamdi Mohammed, A., Alsaif, A., Bokhari Areej, A., Fatima, Qureshi Viquar, Rehan, Qureshi Mohammad, 2013. Review on breast cancer in Kingdom of Saudi Arabia. Middle-East J. Sci. Res. 14 (4), 532–543.

Al-Qahtani, M.S., 2007. Gut metastasis from breast carcinoma. Saudia Med. J. 28, 1590–1592.

American Cancer Society, 2012. Cancer Facts & Figures. American Cancer Society (ACS), Atlanta.

Austin, P.C., Tu, J.V., 2004. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J. Clin. Epidemiol. 57, 1138–1146.

Bagley Steven, C., Halbert, White, Beatrice, Golomb, 2001. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. J. Clin. Epidemiol. 54, 979–985.

Braithwaite, D., Miglioretti, D.L., Zhu, W., Demb, J., Trentham-Dietz, A., Sprague, B., Tice, J.A., Onega, T., Henderson, L.M., Buist, D.S., Ziv, E., 2018. Family history and breast cancer risk among older women in the breast cancer surveillance consortium cohort. JAMA Internal Med. 178 (4), 494–501.

Chang-Claude, J1., Andrieu, N., Rookus, M., Brohet, R., Antoniou, A.C., Peock, S., Davidson, R., Izatt, L., Cole, T., Noguès, C., Luporsi, E., Huiart, L., Hoogerbrugge, N., Van Leeuwen, F.E., Osorio, A., Eyfjord, J., Radice, P., Goldgar, D.E., Easton, D.F., 2007. Epidemiological Study of Familial Breast Cancer (EMBRACE). International BRCA1/2 Carrier Cohort Study (IBCCS) collaborators group 16 (4), 740–746.

Collaborative Group on hormonal factors in breast cancer, 2001. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet. 27; 358 (9291) : 1389-99.

David, Collett, 1991. Modeling binary data. Chapman & Hall/CRC Texts in Statistical Science, London.

Concato, J1., Feinstein, A.R., Holford, T.R., 1993. The risk of determining risk with multivariable models. Ann. Intern. Med. 118, 201–210.

Cox, D.R., Snell, E.J., 1989. Analysis of binary data. Chapman and Hall/CRC, London.

Dall Genevieve Victoria and Britt Kara Louise, 2017. Estrogen effects on the mammary gland in early and late life and breast cancer risk. Frontiers in oncology, 7, 110.

Dawood Shaheenah, S., Xiudong, Lei, Rebecca, Dent, Mainwaring Paul, N., Sudeep, Gupta, Javier, Cortes, Gonzalez-Angulo Ana, M., 2014. Impact of marital status on prognostic outcome of women with breast cancer. J. Clin. Oncol. Breast Cancer—HER2/ER.

Elkum, N., Al-Tweigeri, T., Ajarim, D., Al-Zahrani, A., Amer, S.M., Aboussekhra, A., 2014. Obesity is a significant risk factor for breast cancer in Arab women. BMC Cancer 14, 788.

Genuer, Robin, Poggi, Jean-Michel, Tuleau-Malot, Christine, 2010. Variable selection using random forests. Pattern Recogn. Lett. 31 (14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Andreas, Hadjisavvas, Loizidou Maria, A., Nicos, Middleton, Thalia, Michael, Rena, Papachristoforou, Eleni, Kakouri, Maria, Daniel, Panayiotis, Papadopoulos, Simon, Malas, Yiola, Marcou, Kyriacos, Kyriacou, 2010. An investigation of breast cancer risk factors in Cyprus: a case control study. BMC Cancer 10, 447.

Hopper John, L., Dite Gillian, S., MacInnis Robert, J., Yuyan, Liao, Nur, Zeinomar, Knight, Julia A, Southey, Melissa C., Milne Roger, L., Chung Wendy, K., Giles Graham, G., 2018. Genkinger Jeanine M, McLachlan Sue-Anne, Friedlander Michael L, Antoniou Antonis C, Weideman Prue C, Glendon Gord, Nesci Stephanie, Investigators kConFab. Andrulis Irene L, Buys Saundra S, Daly Mary B, John Esther M, Phillips Kelly Anne and Terry Mary.

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2000. Applied logistic regression. wiley-series in probability and statistics. A Wiley Inter Science Publication, New York.

Loren, Lipworth, Renee, Bailey, Dimitrios, Trichopoulos, 2000. History of breast-feeding in relation to breast cancer risk: a review of the epidemiologic literature. J. Natl. Cancer Inst. 92 (4), 302–312.

Ravichandran, K., Nasser, Al Hamdan, Rahman, Al Dyab Abdul, 2005. Population based survival of female breast cancer cases in Riyadh Region, Saudia Arabia. Asian Pac. J. Cancer Prev. 6, 72–76.

Ravichandran, K., Al-Zahrani, A.S., 2009. Association of reproductive factors with the incidence of breast cancer in Gulf Cooperation Council countries. Eastern Mediterranean Health J. 15 (3), 612–621.

Uddin, S., Ullah, A., Iqbal, M., 2010. Statistical modeling of the incidence of breast cancer in NWFP. In: Pakistan, Journal of applied quantitative methods, pp. 159–165.

Saudia Cancer Registry , 2005, 2008. http://www.scr.org.sa.

Yusuff, H., Mohamad, N., Ngah, U.K., Yahaya, A.S., 2012. Breast cancer analysis using logistic regression. IJRRAS 10 (1), 14–22.