Internal Medicine and Care

Research Article ISSN: 2515-1061

Prospective Breast Cancer Risk Factors Prediction in Sudanese Women

Sawsan Babiker[1]*, Yousif Eltayeb[2] and Bassam Ibrahim[3]

[1]Gezira University, Faculty of Education, Sudan

[2]King Abdulaziz Cardiac Center, King Abdulaziz Medical City - Riyadh, P.O. Box 22490, Riyadh 11426, Saudi Arabia

[3]Head of Statistics &Research Abu Dhabi, UAE

Abstract

Breast cancer is an emerging disease in Sudan. The exact extent is unclear dueto several factors including the lack of a population-based registry. There is, however, a need to conduct basic descriptive studies of cancer. The study aimed to assess the importance of known risk factors for female breast cancer in Sudan using logistic regression models. 100 patient cases for different stages of breast cancer were used tostudy case management, and 100 healthy women from the National Cancer Institute (NCI), Gezira University, Sudan, were taken to predict the probability of women developing breast cancer, A standardized questionnaire was administered to all participants and consisted of socio-demographic factors, obstetric and gynecologic histories, anthropometric measurements, and other variables identified as risk factors from the literature. logistic regression was analyzed by taking factors such as age, marital status, family history, parity, age at first full-term pregnancy, menopausal status, Body Mass Index (BMI), and breastfeeding. 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 breastfeeding) in the development of breast cancer. Findings suggest that the risk factors operative in the development of breast cancer in Sudan are not the same as those identified in more developed nations. Women of lower educational level and early age at menarche with higher body mass index were found to be at significantly increased risk of breast cancer.

Introduction

Cancer in Africa is under-recognized. Public and professional attention has been drawn more to news reports of death due to HIV, infectious diseases, and periodic famines. This is exacerbated by a lack of epidemiologic research and population-based registry data to accurately quantify the problem. Approximately (4%) of all deaths in Africa are attributable to cancer compared to (13%) worldwide[1]. However; the cancer toll across the continent is expected to climb as the population ages and adopts more Westernized behaviors. Breast cancer is the second most common cancer in African females accounting for (19%) of malignancies [1]. Similar proportions (12.9%) have been recorded for the Sudan [2]. In Africa as a whole, breast cancer is less common than cervical cancer; however, it is the most common malignancy in North Africa and certain subpopulations of Sub-Saharan, Africa. For example, in Nigeria, the Ibadan Cancer Registry has now documented that breast cancer is the most common female cancer and the most common cancer among both sexes [3]. Despite increased awareness campaigns about breast cancer, the majority of cancers (80-85%) are still present in advanced stages. Growing awareness of the prevalence of breast cancer in developing countries, as well as the advanced stage of the disease at presentation, has increased the importance of descriptive studies of this disease in the African continent. This manuscript aims to present results from a case-control study conducted at the NCI in Sudan. The importance of this research lies in its ability to investigate whether risk factors known to increase breast cancer in developed countries play a similar role in Sudan. These data may also aid policymakers in the design and implementation of health services to improve the health of women in Sudan.

Materials and Methods

Study design

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

In addition to specialist and pathology records from which risk factors can be identified, the data collection of cases with breast cancer is accomplished by analyzing patient information through a direct interview between the patient and the related clinician.

Case Sample

Cases selected for inclusion in this study were randomly selected from those presenting at NCI with a diagnosis of breast cancer, were aged 25-80 years, had no prior history of breast cancer, and resided in Sudan. All cases were histologically confirmed. Data was collected

**Correspondence to:* Sawsan Babiker, Gezira University, Faculty of Education, Sudan, E-mail: sawsanbabiker@gmail.com

Keywords: *breast cancer, risk factors, logistic regression, sudanese women*

Received: May 08, 2024; **Accepted:** June 03, 2024; **Published:** June 06, 2024

Table 1. Descriptive statistics of cases and controls according to continuous variables

	Cases Mean ± S.E	Control Mean ± S.E.
Age	46.5 ± 1.2	46.4 ± 1.8
Age at menarche	13.2 ± 0.68	13.3 ± 0.72
Age at first full-term pregnancy	22.3 ± 1.6	22.1 ± 1.4
Age at menopause	44.9 ± 0.76	45.1 ± 0.69
Body Mass Index (BMI)	25.2 ± 1.3	24.7 ± 1.3

Table 2. Distribution of cases and control according to age groups.

Variables	Case No (%)	Control No (%)	χ^2	P -value
Age group (years old)			0.531	0.971
20 - 30	4 (4.0)	4 (4.0)		
31– 40	37 (37.0)	38 (38.0)		
41–50	31 (31.0)	28 (28.0)		
51– 60	16 (16.0)	15 (15.0)		
Above 60	12 (12.0)	15 (15.0)		

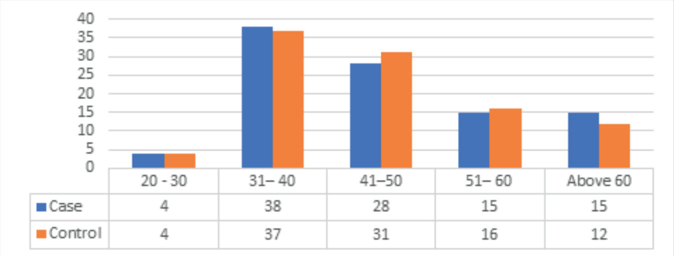

Figure 1. Distribution of cases and control according to age groups

through a 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). The diagnosis of cases with breast cancer was the response factor for the study and from the patient's direct interview, the missing information was completed.

In addition to specialist and pathology records from which risk factors can be identified, the data collection of cases of breast cancer is accomplished by analyzing patientinformation through a direct interview between the patient and the related clinician.

Control Sample

The control women were recruited randomly, residing in the same geographical region, and admitted to the NCI without a history of breast problems or neoplastic diseases and who resided in the same geographical region as the case. Control cases were matched for gender and age; women confirmed diagnosis with breast cancer were interviewed by one investigator.

Data Set

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

This study was conducted based on a sample of 200 people, including 100 cases (cases with breast cancer) and 100 control cases (not cases with breast cancer). Among women with breast cancer, 92 (92.0%) and 81 (81.0%) 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.

Methods

We have followed Salah, et al. methods. The relationship between a binary variable and one or more explanatory values is defined by the logistic regression method (Appendix -1) according to [1,4-8].

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 about breast cancer risk to calculate odds ratio (OR) and 95% Confidence Intervals (CI), as illustrated by Tables of the (Appendix-1), equations 1 to 4 of (Appendix-2) [9,10] andEquation 5 of (Appendix-3), [9-15]. Risk factors associated with breast cancerhave been entered into a multivariate logistic regression analysis of the forward-looking range (Appendix-4).

Results

Socio-demographic factors

Age

Breast cancer cases and controls were detected in cases as young as 23 years and asold as 81 years with a mean ± S.E. 46.5 ± 1.2 and 46.4 ± 1.8 years for cases and controls, respectively as shown in (Table 1).

Results from (Table 2), (Figure 1) show that the maximum risk factors are in the age group of 31 - 40 with the cases of 37 out of 38 control samples, followed by 31 cases of breast cancer from the age group of 41 - 50 out of 28 controls and less case of 4 out of 4 was observed in the age group with less than 20 - 30.

Marital status

However, results from (Table 3), (Figure 2) observed in married cases were high with 81 cases out of 80, compared to 4 cases with divorced out of 3, and 7 cases and 5 of control are widowed and 4 cases

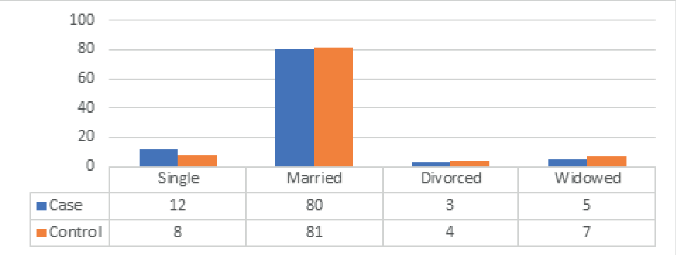

Figure 2. Distribution of cases and controls according to marital status

Table 3. Distribution of cases and control according to marital status

Variables	Case No (%)	Control No (%)	χ^2	P -value
Marital status			1.282	0.733
Single	8 (8.0)	12 (12.0)		
Married	81 (81.0)	80 (80.0)		
Divorced	4 (4.0)	3 (3.0)		
Widowed	7 (7.0)	5 (5.0)		

Table 4. Distribution of cases and controls according to Body Mass Index (BMI)

Variables	Case No (%)	Control No (%)	χ^2	P -value
BMI			3.996	0.182
< 20	21	19		
20 – 24	30	27		
25 – 29	26	23		
30 - 34	16	15		
>= 35	7	16		

Table 5. Distribution of cases and controls according to Education level

Variables	Case No (%)	Control No (%)	χ^2	P -value
	Education level		5.190	1.58
Illiterate	47	42		
Primary	37	29		
Secondary	11	18		
University	5	11		

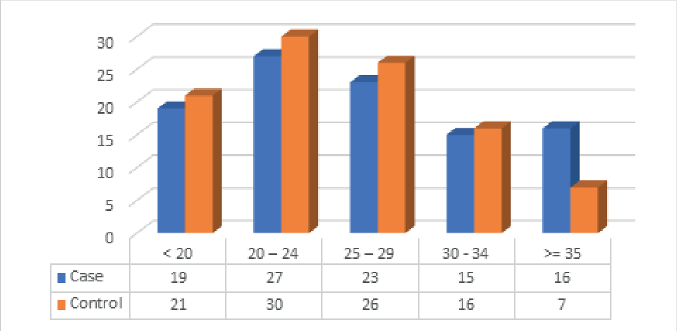

Figure 3. Distribution of cases and controls according to BMI

and 3 of control are divorced, and 8 of cases and 12 of control are singles.

Body Mass Index (BMI)

As shown in (Table 4), (Figure 3), BMI had a significance p-value (of 0.000) in which (49) Of the cases were obese, whereas 54% of control subjects were obese (Figure 3). More cases were observed with a BMI of 20 - 24 with cases of 30 out of 27, 26 cases out of 23 controls were observed with a BMI of 25-29, followed by the cases with 21 out of 19 with aBMI less than 20, and 16 out of 15 with BMI 30 –34, and at BMI more than 35 there are 7 0ut of 16.

Education Level

Education is known to have important effects on all aspects of human life. (Table 5), (Figure 4) gives the distribution of the cases and control according to education level. Most of the cases are illiterate (47%, 42%) for cases and controls respectively, whereas (37%, 29%)for cases and control respectively are primary, (11%,18%) of cases and control are secondary and (5%,11%) are university. The difference between the distributions of cases and control concerning educational level is statistically significant at the 5% level, with p-value = 0.000.

Age at Menarche

Information regarding age at menarche was available, mean age at menarche was found to be 13.2 ± 0.68 years for cases and 13.3 ± 0.72 years for control. This difference between the mean ages at menarche was statistically not significant (p-value = 0.647).(Table 6), (Figure 5) illustrates that most of the cases had menarche >12 years (88). The conclusion drawn from these results is that most of the Sudanese female, have their menarche between ages 10 and 14 with a mean of age 13. Menarche at advanced age isvery rare even between both study groups.

Parity

Results from (Table 7), (Figure 6), show that the maximum risk factors are in the parityof more than 4 children with cases of 39 out of 31 in controls, followed by 33 cases from never conceived (nulliparous) out of 35 controls and less case of 12 out of 16 was

Table 6. Distribution of cases and control according to age at menarche

Variables	Case No (%)	Control No (%)	χ^2	P -value
	Age at menarche		0.831	0.50
≤ 12 years	12	13		
> 12 years	88	87		

Table 7. Distribution of cases and control according to parity

Variables	Case No (%)	Control No (%)	χ^2	P -value
	Parity		1.662	0.645
Nullpariuos	33	35		
1 – 2 child	16	18		
3 – 4 child	12	16		
> 4 Child	39	31		

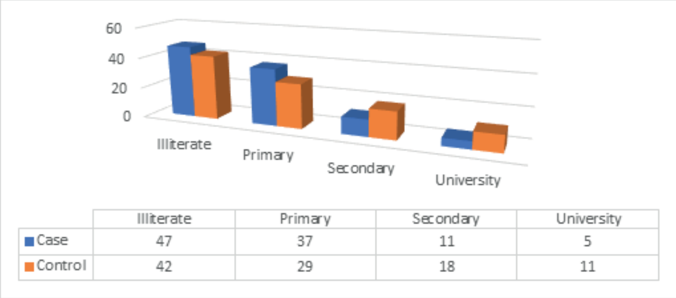

Figure 4. Distribution of cases and controls according to Education level

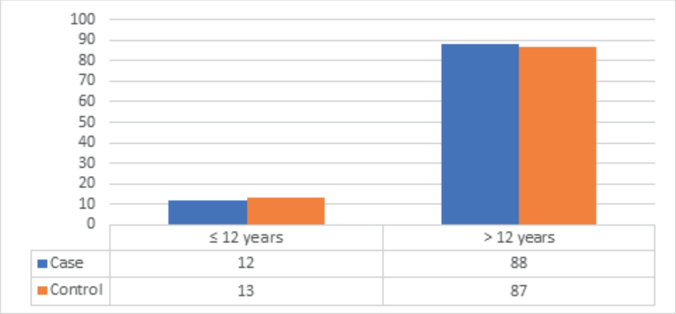

Figure 5. Distribution of cases and controls according to Age at menarche

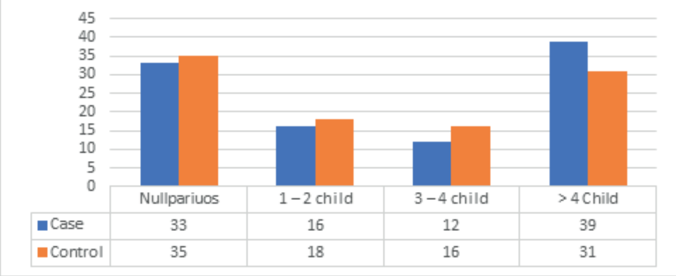

Figure 6. Distribution of cases and controls according to Parity

Table 8. Distribution of cases and control according to age at 1st full-term pregnancy

Variables	Case No (%)	Control No (%)	χ^2	P -value
Age at 1st full-term pregnancy				
Ever	33	35	0.784	0.941
< 20 year	25	21		
20 – 24 year	21	21		
25 – 29 year	13	16		
≥ 30 year	8	7		

Table 9. Distribution of cases and controls according to menopausal status

Variables	Case No (%)	Control No (%)	χ^2	P -value
menopausal status				
post-menopause	47	36	2.492	0.054
pre-menopause	53	64		

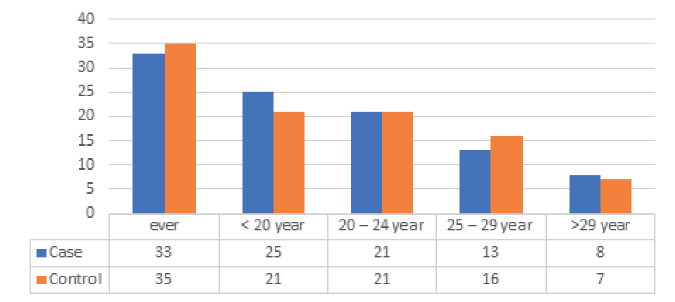

Figure 7. Distribution of cases and controls according to **age at 1st birth of child**

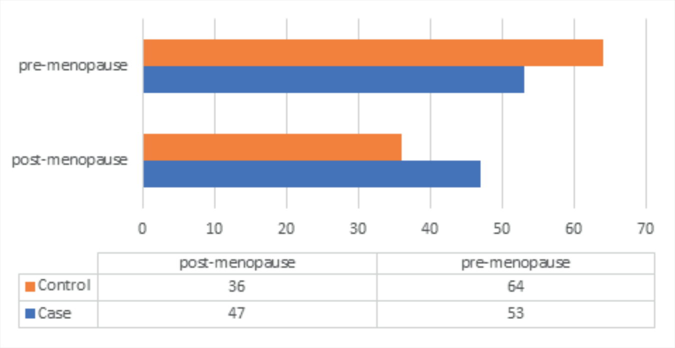

Figure 8. Distribution of cases and controls according to menopausal status

observed in the parity of 1 – 2 children. The difference between the distributions of cases and control about parity is statistically insignificant, with p-value = 0.645

Age at Birth of first child

Age at first full-term pregnancy as shown in (Table 8), (Figure 7) ranged from 15 – 41 years among control. The mean age at first full-term pregnancy was 22.3 ± - 4.6 years among cases and 22.1 ± 3.4 years among control. Age at 1st full term pregnancy about (33%) of cases and (35%) of control are at the range of ever have child, (25%) of cases and (21%) of control their age at first full-term pregnancy between the age groups less than 20 years, (21%) of cases and (21%) of control their age 20 – 24 years, 13% of casesand 16% of control are 25 – 29 years and 8% of cases and 7% of control at the age ≥ 30.

Menopausal Status

The mean age at menopause was 44.9 ± 0.76 years for cases and 45.1 ± 0.69 years for control. The difference between the two means was statistically significant ata 5% level (p-value < 0.01). (Table 9), (Figure 8) illustrates that, at presentation 53%and, 64% of cases and control respectively, were premenopausal, and 47% and, 36% of cases and controls respectively were postmenopausal. This indicates that premenopausal women are at higher risk for developing breast cancer. The data from developing countries shows a high incidence of breast cancer among postmenopausal women rather than in premenopausal women. This difference between these two data may reflect the different natural history of the disease in developing countries, short life expectancy, or maybe it is a result of many other factors that are starting to be explained. The difference between the distributions of cases and control about menopausal status is statistically significant, with p-value = 0.054.

Contraceptive Use

Distribution of cases and controls according to contraceptive use as shown in (Table 10), (Figure 9), illustrates that most cases have never used contraception in their life (78%). Only (22%) had used it, (62%) of control never used it and (38%) used it. The difference between the distributions of cases and control about contraceptive use is statistically significant, with p-value = 0.01.

Residence

The distribution of cases and controls according to residence is shown in (Table 11), and (Figure 10). Illustrate that most of the cases

Table 10. Distribution of cases and controls according to contraceptive use

Variables	Case No (%)	Control No (%)	χ^2	P -value
contraceptive use				
Yes	22	38	6.095	0.01
No	78	62		

Table 11. Distribution of cases and controls according to residence

Variables	Case No (%)	Control No (%)	χ^2	P -value
residence				
Rural	69	68	0.023	0.50
Urban	31	32		

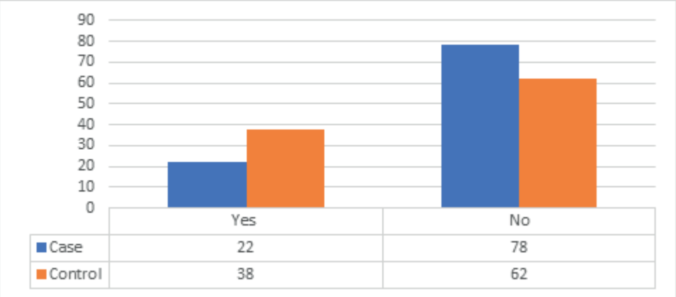

Figure 9. Distribution of cases and controls according to contraceptive use

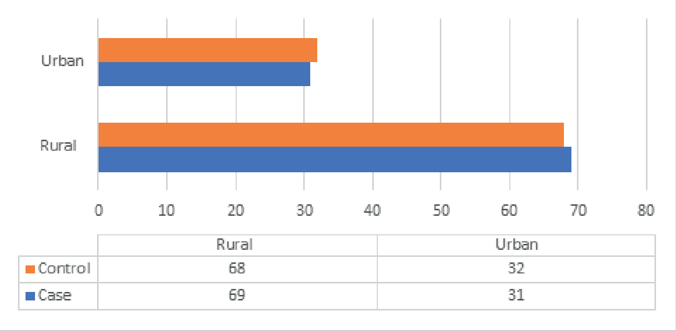

Figure 10. Distribution of cases and controls according to residence

Table 12. Distribution of cases and controls according to breastfeeding

Variables	Case No (%)	Control No (%)	χ^2	P -value
breast feeding			0.347	0.329
Yes	66	62		
No	34	38		

Table 13. Distribution of cases and controls according to HRT

Variables	Case No (%)	Control No (%)	χ^2	P -value
	hormonal replacement therapy		0.189	0.414
Yes	11	13		
No	89	87		

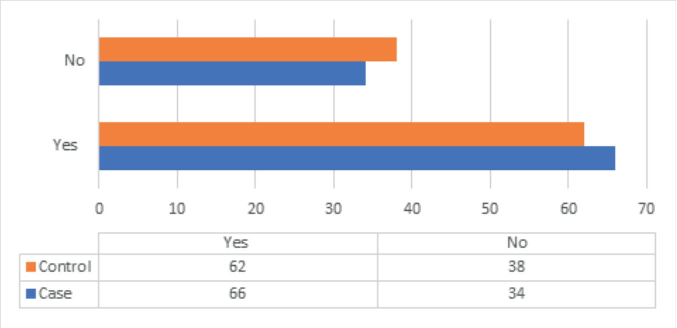

Figure 11. Distribution of cases and controls according to breastfeeding

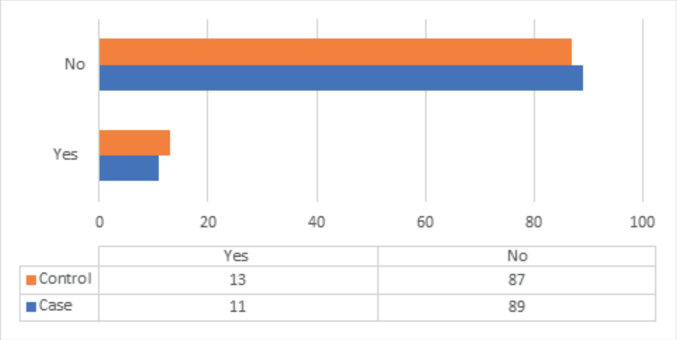

Figure 12. Distribution of cases and controls according to (HRT)

and control came from rural areas (69%, and 68%) respectively, whereas (31%, and 32%) respectively came from urban areas. The Census of 1993 showed that most of the population lived in rural areas and 29% of Sudan's population lived in urban areas, whereas, less than 3% of the populations were nomads [16]. The chi-square test suggests that the difference in the distributions of rural and urban breast cancer is statistically insignificant, with a p-value = 0.50.

Breast Feeding

The distribution of cases and controls according to breastfeeding as shown in (Table 12) and (Figure 11) illustrates that most cases and controls had breastfeeding (66%, 62%) respectively, compared with women who hadn't (34%, 38%). The difference between the distributions of cases and control about breastfeeding is statistically insignificant, with p-value = 0.329.

Hormonal Replacement Therapy (HRT)

The distribution of cases and controls according to HRT is shown in (Table 13) and (Figure 12). The case study shows that 89% of cases were not treated by HRT, 11% did that, 87% of the control group weren't treated with HRT, and 13% did it. The difference between the distributions of cases and control about HRT is statistically insignificant, with p-value = 0414

Previous Benign Biopsy (PBB)

The distribution of cases and controls according to PBB is shown in (Table 14) and (Figure 13). According to the case study shows that 77% of cases didn't have PBB, 23% had that, 84% of control didn't have a previous benign biopsy, and 16% had it. The difference between the distributions of cases and control about PBB is statistically insignificant, with p-value = 0.142.

Occupation

The distribution of cases and controls according to occupation is shown in (Table 15) and (Figure 14). According to the case study were observed with cases 86% are housewives, 14% are employees, 67% of control is housewives, and 33% are employees. The difference between the distributions of cases and control about occupation is statistically significant, with p-value = 0.003.

Table 14. Distribution of cases and controls according to PBB

Variables	Case No (%)	Control No (%)	χ^2	P -value
previous benign biopsy (PBB)			1.561	0.142
Yes	23	16		
No	77	84		

Table 15. Distribution of cases and controls according to occupation

Variables	Case No (%)	Control No (%)	χ^2	P -value
occupation			8.521	0.003
Housewife	86	67		
Employee	14	33		

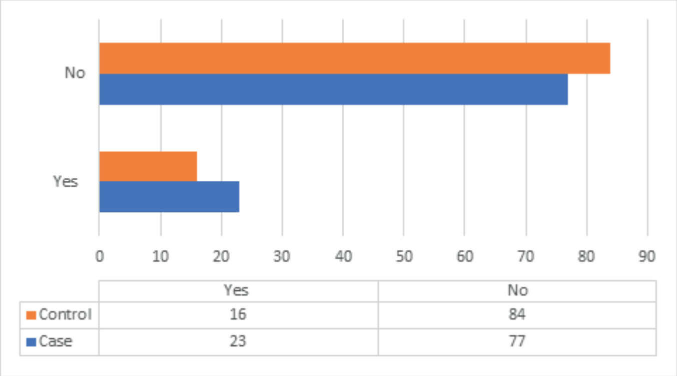

Figure 13. Distribution of cases and controls according to PBB

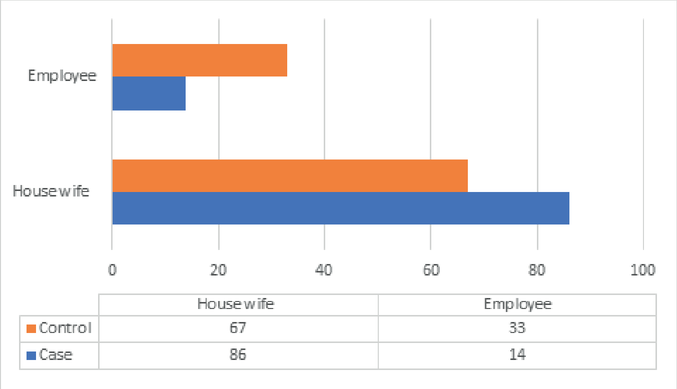

Figure 14. Distribution of cases and controls according to occupation

Table 16. Distribution of cases and controls according to family history of breast cancer

Variables	Case No (%)	Control No (%)	χ^2	P -value
family history of breast cancer			1.705	0.138
Yes	9	15		
No	91	85		

Table 17. Distribution of cases and controls according to tumor stage

Tumor stage	Frequency N (%)
I	3 (3%)
IIA	11 (11%)
IIB	17 (17%)
IIIA	19 (19%)
IIIB	20 (20%)
IIIC	5 (5%)
IV	24 (24%)
recurrence	1 (1%)
Total	100 (100%)

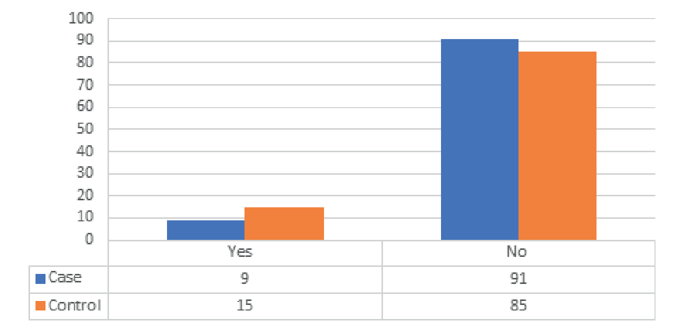

Figure 15. Distribution of cases and controls according to family history of breast cancer

Family history of breast cancer

Distribution of cases and controls according to family history of breast cancer is shown in (Table 16) and (Figure 15) in the case study the women with a history of breast cancer were observed with cases 9 %, and 91% haven't a family history of breast cancer, 15% of control is from a family with history of breast cancer while 85 % observed that they have not family history of breast cancer. The difference between the distributions of cases and control about family history of breast cancer is statistically insignificant, with p-value = 0.138

Tumor Stage

Breast cancer is a malignant tumor of breast tissue suspected in individuals with clinical findings such as a breast lump, breast thickening skin change, or changes on a mammogram. Breast cancer is staged from 0 to IV, with survival dependent upon the stage at diagnosis as we see in (Table 17) and (Figure 16) 24% of cases in stage IV, 20% atstage IIIB, and 1% recurrence, so we need to promote breast cancer education and awareness among the public policymakers, health professional and the media and population.

All variables show significant variation, (Table 18.a), by using Model -1 as follows:

$$logit\ \hat{p} = 3.678 + (0.546)\ age - (0.776)\ family\ history + (0.616)\ contraceptive\ use +$$

$$(0.892)\ occupation + (1.246)\ education\ level - (1.182)\ age\ at\ menarache - (0.495)\ residence + (1.5)\ BMI \qquad (1)$$

So, we see in tables (18.b.1, 18.b.2), the model with significant covariate, that the chi-square statistic for the likelihood ratio tests, where pr ($\chi^2 \geq 73.73$) = 0.00 with 8 d. f. is highly significant, from table (18.b.3) the p-value of Pearson goodness of fit test equals 0.285 and the p-value of Deviance goodness of fit test is 0.277, this means that the mode is l well fitting. Therefore, the risk factors of breast cancer among the study group are due to these variables: age, family history of breast cancer, contraceptive use, occupation, education level, age at

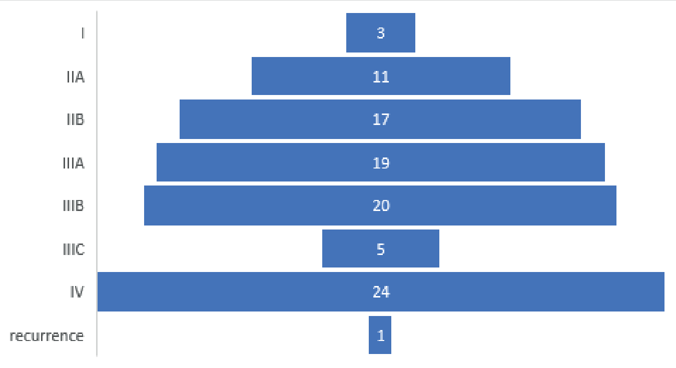

Figure 16. Distribution of cases and controls according to Tumor stage

Table 18.a. Estimated coefficients for variable in Model -1

Variable	B	S. E.	Wald	Sig.	$\hat{OR}$	95% CI for OR
Age Group	0.546	0.236	5.335	0.021	1.726	1.086, 2.743
Family History of Breast Cancer	-0.776	0.462	2.825	0.093	0.460	0.186, 1.138
Contraceptive Use	0.616	0.366	2.829	0.093	0.540	0.264, 1.107
Occupation	0.891	0.414	3.570	0.050	2.186	0.122, 4.922
Education Level	1.246	0.195	40.665	0.000	3.477	2.371, 5.100
Age at Menarche	-1.182	0.423	4.435	0.035	2.437	1.064 , 5.583
Residence	-0.495	0.372	1.768	0.184	0.610	0.294, 1.264
BMI	0.150	0.139	1.173	0.279	1.162	0.886, 1.524
Constant	3.658	1.319	7.696	1	0.006	

Table 18.b.1 Model assessment

Model	-2 Log Likelihood	Chi-Square	d.f.	Sig.
Null	251.599			
Final	177.869	73.730	8	0.000

Table 18.b.2 Model assessment

Effect	-2 Log Likelihood	Chi-Square	Sig.
Age group	183.530	5.661	0.017
Family History of Breast Cancer	180.757	2.888	0.089
Contraceptive Use	180.753	2.884	0.089
Occupation	182.474	4.605	0.032
Education Level	236.361	58.492	0.000
Age at Menarche	185.795	7.926	0.005
Residence	179.674	1.805	0.179
BMI	179.055	1.186	0.276

Table 18.b.3 Model assessment Goodness of fit test

	Chi-Square	Sig.
Pearson Deviance	156.285 156.708	0.285 0.277

menarche, residence, and BMI. All variables show significant variation, (Table 19.a), by model -2 as follows:

$Logit\ (\hat{p})$ = 2.249 + 0.498 age group – 0.903 family history ++ 1.203 education level – 1.1079 age at menarche + 0.234BMI (2)

Finally, we see in tables (19.b.1, 19.b.2), the final model with significant covariate, that thechi-square statistic for the likelihood ratio tests, where pr ($\chi^2 \geq 64.517$) = 0.00 with 4 d. f. is highly significant, from table (19.b.3) the p-value of Pearson goodness offit test equals 0.813 and the p-value of Deviance goodness of fit test is 0.703, this means that the model well fitting. Therefore, the risk factor of breast cancer amongthe study group due to these variables: age, family history of breast cancer,education level, and age at menarche.

The evaluation of the Model in (Table 19.b.4), showed that R^2 = 0.77825 and the adjusted R^2 is 0.73390, in addition, the R^2 value was good and showed statistically significant forecasts (P-value < 0.05). Important assumptions were made about the relationship between changes in predictor values and changes in response values. 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.

Discussion

Backward elimination was conducted using SPSS version 22 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 BMI and an increase in the number of cases of breast cancer (Hopper, 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 [17]. Inaddition, 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 [4,17]. Family history is a risk factor in previous studies [18], the logistic regression model is one of the best models used to determine risk factors [19].

Table 19.a. Estimated coefficients for variable in Model -2

Variable	B	S. E.	Wald	Sig.	$\hat{OR}$	95% CI for $\hat{OR}$
Age group	0.498	.216	5.295	0.021	1.646	1.077, 2.516
Family History of Breast Cancer	- 0.903	0.393	5.273	0.022	0.405	0.188, 0.876
Education Level	1.203	0.179	45.075	0.000	3.329	2.343, 4.729
Age at Menarche	-1.079	0.394	7.488	0.006	0.340	0.157 , 0.736
BMI	0.234	0.127	3.173	0.012	1.562	1.026, 1.952

* Reduce some variables: contraceptive use, occupation, and residence

Table 19.b.1 Model assessment

Model	-2 Log Likelihood	Chi-Square	d.f.	Sig.
Null Final	128.987 64.470	64.517	4	0.000

Table 19.b.2 Model assessment

Effect	-2 Log Likelihood	Chi-Square	Sig.
Age group	70.070	5.600	0.018
Family History of Breast Cancer	70.069	5.599	0.018
Education Level	128.478	64.008	0.000
Age at Menarche	72.328	7.585	0.005
BMI	72.055	6.186	0.002

Table 19.b.3 Model assessment Goodness of fit test

	Chi-Square	Sig.
Pearson Deviance	22.183 24.524	0.813 0.703

Table 19.b.4 Model Summary

Last model assessment	
Multiple R	0.88218
R Square	0.77825
Adjusted R Square	0.73390
Standard Error	1.11437

Table 20. Estimated coefficients for a multiple logistic regression model (2)

	B	S.E.	Wald	Sig.	$\hat{OR}$	95% C.I. for $\hat{OR}$	
						Lower	Upper
Age group			0.682	.953			
20 - 30	0.457	0.839	0.296	0.587	1.579	0.305	8.182
31– 40	0.240	0.475	0.254	0.614	1.271	0.501	3.226
41–50	0.376	0.490	0.590	0.443	1.456	0.558	3.803
51– 60	0.199	0.552	0.130	0.718	1.221	0.413	3.604
Education level			5.297	0.151			
Illiterate	1.021	0.615	2.752	0.097	2.775	0.831	9.268
Primary	1.038	0.624	2.764	0.096	2.823	0.830	9.599
Secondary	0.283	0.691	0.167	0.683	1.326	0.343	5.137
Breastfeeding (no)	0.349	0.315	1.224	0.269	1.417	0.764	2.629
BMI			3.283	0.512			
20 – 24	0.891	0.570	2.439	0.118	2.437	0.797	7.450
25 – 29	0.828	0.542	2.936	0.087	2.531	0.875	7.319
30 - 34	0.848	0.552	2.354	0.125	2.334	0.790	6.894
≥ 35	0.943	0.592	2.028	0.154	2.532	0.728	7.420
Family history of BC (yes)	-0.739	0.471	2.466	0.116	0.478	0.190	1.201
Age at Menarche ≤ 12 years	-1.052	0.4787	4.8248	0.0281	0.3494	0.1367	0.8929

In the current study, breastfeeding did not play a protective role in breast cancer, since a smaller number of breastfeeding cases were observed. Some studies suggest it is possible to prevent breast cancer by breastfeeding and some studies have shown that breastcancer risk does not affect lactation [20]. Nevertheless, epidemiological studies have indicated that populations with normal long lactation periods pose low breast cancer risks [20]. These conflicting results suggest that the effects of breast cancer risk factors are likelyto 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 lowering breast cancer [21], menopausal stages affect the risk of breast cancer [22-24].

Conclusion

Based on our data and tables suggested that the risk factor for developing breast cancer was in theage group of 30 – 30, those who are married have a BMI ≥ 30, bear fewer children, not breastfeeding, 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 fewer cases of breast cancer, data also suggest us that the women bearing children >10 and also breastfeeding plays as a protective role in developing breast cancer, and also less number of cases were observed with menopausal status at the age > 45 (Table 20).

Acknowledgments

This study was supported by Professor Arthur M. Michalek, Professor of Oncology and Epidemiology, Senior Vice President, and Dean of Educational Affairs, at Roswell Park Cancer Institute, USA. Professor Bassam Ibrahim, Head of Applied Statistics Dept., College of Science, Sudan University of Science & Technology, Sudan.

Declaration of Competing Interest

Author(s) declare that all works are original and this manuscript has not been published inany 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 standards are open to bias. Recall bias was alsoexpected as regards their date e.g. age, age of 1st pregnancy, number of children, breastfeeding.

Author Contributions Conceptualization

Sawsan Babiker

Data curation

Sawsan Babiker

Formal analysis

Sawsan Babiker

Methodology

Sawsan Babiker

Resources

National Cancer Institute, Gezira University, Medani, Sudan.

Writing – original draft

Sawsan Babiker and Yousif Eltayeb

Writing – review & editing

Sawsan Babiker and Yousif Eltayeb

Appendices

Appendix -1

Methods

We followed the methods of [8]. 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 [9,10,17].

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

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

$$g(x) = \log t\ (p(x)) = \mathrm{h}\ \frac{p(x_i)}{1 - p(x_i)} = \beta_0 + \beta_i x \qquad (2)$$

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

The parameters in this model can no longer be estimated by least squares, but are found using themaximum 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:

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

This value is distributed as a 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− α) % 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:

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

This log transformation of the likelihood functions yields a chi-squared statistic. These are 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 entireor 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 [15].

Appendix-2

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 the main data. First, thedata were divided into two groups. 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 value, 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 are 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 [5].

Appendix-3

Variable selection

It is critical that the model contains all relevant variables and does not start with more than the number of observations justified [9,10,13]. Additional variables typically produce a better model that fits the data for a dataset. Excessive variables, however, influence the model coefficient and help over fit 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 [12,14]. 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 [4]. The study of co-linearity wasbased on essential interaction test values. Each variable must have significant values less than 0.20 [12], used in the study of the logistic regression model [5].

Table 21.a. Distribution of age groups according to age at 1st pregnancy

			Age at 1st pregnancy					Total	**p-value**
			ever	**< 20 year**	20 - 24 year	25 - 29 year	**> 29 year**		
control	**age group**	20 - 30	2	1	1	0	0	4	0.045
		31 - 40	11	8	7	6	6	38	
		41 - 50	7	6	9	5	1	28	
		51 - 60	6	3	2	4	0	15	
		60 and above	9	3	2	1	0	15	
	Total		35	21	**21**	**16**	**7**	**100**	
cases	**age group**	20 - 30	2	0	2	0	0	4	0.015
		31 - 40	16	10	6	3	2	37	
		41 - 50	11	5	6	6	3	31	
		51 - 60	3	7	4	1	1	16	
		60 and above	1	3	3	3	2	12	
		Total	**33**	**25**	**21**	**13**	**8**	**100**	

Table 21.b. Distribution of Age groups according to body mass index BMI

			body mass index BMI						
			< 20	20 - 24	25 - 29	3 - 34	≥ 35	Total	p-value
control	**age group**	20 - 30	1	1	0	1	1	4	0.419
		31 - 40	6	11	9	6	6	38	
		41 - 50	8	9	5	3	3	28	
		51 - 60	1	2	5	3	4	15	
		60 and above	3	4	4	2	2	15	
	Total		19	27	23	15	16	100	
cases	**age group**	20 - 30	0	4	0	0	0	4	0.314
		31 - 40	11	8	11	6	1	37	
		41 - 50	3	11	6	7	4	31	
		51 - 60	4	3	6	2	1	16	
		60 and above	3	4	3	1	1	12	
	Total		21	30	26	16	7	100	

Table 21.c. Distribution of age groups according to menopausal status

			menopausal status		**Total**	**p-value**
			postmenopause	**pre-menopause**		
control	**age group**	20 - 30	3	1	4	0.136
		31 - 40	13	25	38	
		41 - 50	12	16	28	
		51 - 60	5	10	15	
		60 and above	3	12	15	
	Total		**36**	**64**	**100**	
cases	**age group**	20 - 30	1	3	4	0.000
		31 - 40	4	33	37	
		41 - 50	16	15	31	
		51 - 60	14	2	16	
		60 and above	12	0	12	
	Total		**47**	**53**	**100**	

Table 21.d. Distribution of age groups according to marital status

			Marital Status				**Total**	**p-value**
			married	**widowed**	**divorced**	**single**		
control	**age group**	**20 - 30**	3	0	0	1	4	0.406
		31 - 40	28	2	3	5	38	
		41 - 50	26	0	0	2	28	
		51 - 60	12	1	0	2	15	
		60 and above	11	2	0	2	15	
	Total		**80**	**5**	**3**	**12**	**100**	
cases	**age group**	20 - 30	4	0	0	0	4	0.117
		31 - 40	28	1	3	5	37	
		41 - 50	27	1	0	3	31	
		51 - 60	12	3	1	0	16	
		60 and above	10	2	0	0	12	
	Total		**81**	**7**	**4**	**8**	**100**	

Table 21.e. Distribution of age groups according to breastfeeding

			breast feeding		**Total**	**p-value**
			Yes	**No**		
control	**age group**	20 - 30	2	2	4	0.429
		31 - 40	24	14	38	
		41 - 50	20	8	28	
		51 - 60	9	6	15	
		60 and above	7	8	15	
	Total		62	38	100	
cases	**age group**	20 - 30	2	2	4	0.005
		31 - 40	20	17	37	
		41 - 50	20	11	31	
		51 - 60	13	3	16	
		60 and above	11	1	12	
	Total		66	34	100	

Table 21.f. Distribution of age groups according to number of children

			number of children				Total	p-value
			null	1 - 4	5 - 10	> 10		
control	age group	**20 - 30**	2	0	0	2	4	0.013
		31 - 40	11	8	9	10	38	
		41 - 50	7	9	3	9	28	
		51 - 60	6	1	2	6	15	
		60 and above	9	0	2	4	15	
	Total		35	18	16	31	100	
cases	age group	20 - 30	3	1	0	2	6	0.139
		31 - 40	9	6	8	8	31	
		41 - 50	8	4	3	9	24	
		51 - 60	6	3	2	7	18	
		60 and above	7	5	4	5	21	
	Total		33	19	17	31	100	

Table 21.g. Distribution of age groups according to family history of breast cancer BC

			family history of breast cancer		Total	p-value
			Yes	No		
control	age group	20 - 30	1	3	4	0.484
		31 - 40	6	32	38	
		41 - 50	5	23	28	
		51 - 60	1	14	15	
		60 and above	2	13	15	
	Total		15	85	100	
cases	age group	20 - 30	0	4	4	0.284
		31 - 40	3	34	37	
		41 - 50	4	27	31	
		51 - 60	0	16	16	
		60 and above	2	10	12	
	Total		9	91	100	

Appendix 4

Distribution of Age groups (cases and control) according to risk factors. Table 21.a,21.b,21.c,21.d,21.e,21.f,21.g

References

1. Al DA, Qureshi S, Al Saleh KA, Al Qahtani FH, Aleem A, et al. (2013) Review on breast cancer in Kingdom of Saudi Arabia. *Middle-East Journal of Scientific Research* 14: 532-543.

2. Al-Qahtani MS (2007) Gut metastasis from breast carcinoma. *Saudia Medical Journal* 28: 1590-1592. [Crossref]

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

4. 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* 358: 1389-1399. [Crossref]

5. Concato J, Feinstein AR, Holford TR (1993) The risk of determining risk with multivariable models. *Ann Intern Med* 118: 201-210. [Crossref]

6. Cox DR, Snell EJ (1989) Analysis of Binary Data. 2nd Edition, Chapman and Hall/CRC, London.

7. Ravichandran K, Al Hamdan Nasser, Al Dyab Abdul Rahman (2005) Population-based survival of female breast cancer cases in Riyadh Region, Saudia Arabia. *Asian Pac J Cancer Prev* 6: 72-76. [Crossref]

8. Salah U, Arif U, Najma, Muhammad I (2010) Statistical modeling of the incidence of breast cancer in NWFP, Pakistan. *Journal of applied quantitative methods* 5: 159-165.

9. Austin PC, Tu JV (2004) Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* 57: 1138-1146. [Crossref]

10. Hadjisavvas A, Loizidou MA, Middleton N, Michael T, Papachristoforou R, et al. (2010) An investigation of breast cancer risk factors in Cyprus: a case-control study. *BMC Cancer* 10: 447. [Crossref]

11. Collett D (1991) Modeling Binary Data, Chapman & Hall/CRC Texts in Statistical Science, 2nd Edition, London.

12. Hosmer DW, Lemeshow S (2000) Applied Logistic Regression. Wiley-Series in Probability and Statistics, A Wiley Inter Science Publication, New York.

13. Bagley SC, White H, Golomb BA (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. [Crossref]

14. Genuer R, Poggi, Jean-Michel, Tuleau-Malot C (2009) Variable selection using random forests. *Pattern Recognit Lett* 31: 14.

15. Yusuff H, Mohamad N, Ngah UK, Yahaya AS (2012) Breast cancer analysis using logistic regression. *IJRRAS* 10: 14-22.

16. Sudan Household Health Survey (SHHS) 2nd Round 2010 Summary Report July: Federal Ministry of Health, Ministry of Health, Government of South Sudan, Central Bureau of Statistics Southern Sudan Commission of Census, Statistics & Evaluation.

17. Elkum N, Al-Tweigeri T, Ajarim D, Al-Zahrani A, Amer SMB, et al. (2014) Obesity is a significant risk factor for breast cancer in Arab women. *BMC Cancer* 14: 788. [Crossref]

18. Braithwaite D, Miglioretti DL, Zhu W, Demb J, Trentham-Dietz A, et al. (2018) Family history and breast cancer risk among older women in the breast cancer surveillance consortium cohort. *JAMA Intern Med* 178: 494-501. [Crossref]

19. Dawood SS, Lei X, Dent R, Mainwaring PN, Gupta S, et al. (2014) Impact of marital status on prognostic outcome of women with breast cancer. Journal of Clinical Oncology Breast Cancer-HER2/ER.

20. Lipworth L, Bailey R, Trichopoulos D (2000) History of breast- feeding about breast cancer risk: a review of the epidemiologic literature. *J Natl Cancer Inst* 92: 302-312. [Crossref]

21. Dall GV, Britt KL (2017) Estrogen effects on the mammary gland in early and late life and breast cancer risk. *Front Oncol* 7: 110. [Crossref]

22. Chang-Claude J, Andrieu N, Rookus M, Brohet R, Antoniou AC, et al. (2007) Epidemiological Study of Familial Breast Cancer (EMBRACE). *International BRCA1/2 Carrier Cohort Study (IBCCS) collaborators group* 16: 740-746.

23. Hopper JL, Dite GS, MacInnis RJ, Liao Y, Zeinomar N, et al. (2018) Age- specific breast cancer risk by body mass index and familial risk: prospective family study cohort (ProF-SC). *Breast Cancer Res* 20: 132. [Crossref]

24. Ravichandran K, Al-Zahrani AS (2009) Association of reproductive factors with the incidence of breast cancer in Gulf Cooperation Council countries. *East Mediterr Health J* 15: 612-621. [Crossref]