Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2020;6(4):259-266

Original Article

Evaluating Related Factors to the Number of Involved Lymph Nodes in Patients with Breast Cancer Using Zero-Inflated Negative Binomial Regression Model

Shima Younespour[1], Elham Maraghi[2*], Amal Saki Malehi[2], Maedeh Raissizadeh[2], Mohammad Seghatoleslami[3], Mehran Hosseinzadeh[4]

[1]Dentistry Research Institute, Tehran University of Medical Sciences, Tehran, Iran.
[2]Department of Biostatistics and Epidemiology, Faculty of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.
[3]Environmental and Petroleum Pollutants Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.
[4]Thalassemia and Hemoglobinopathy Research Center, Health Research Institute, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

ARTICLE INFO

Received 05.10.2020
Revised 23.10.2020
Accepted 20.11.2020
Published 10.12.2020

Key words:
Number of involved lymph nodes;
Breast cancer;
Count outcome

ABSTRACT

Background and aims: In Iran, breast cancer accounts for 24.4% of all cancers and contributes to 14.2% of cancer-associated mortality in women. A major challenge facing the health system is to examine the health status of patients with breast cancer, which often involves the axillary lymph nodes. The number of involved nodes should be clinically predicted to ascertain postoperative radiotherapy and chemotherapy. The present study employed regression models to investigate the determinants of the number of lymph nodes involved in patients with breast cancer.

Methods: This retrospective study recruited patients diagnosed with breast cancer during 2005-2015 referring to Shafa Hospital affiliated to Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. The outcome variable was the number of involved lymph nodes. Regression models for count outcomes, were utilized for investigating the related factors to the number of involved lymph nodes in patients with breast cancer.

Results: A sample of 165 patients was eligible for the present study. The Akaike information criterion (AIC) of the zero-inflated negative binomial (ZINB) model was the lowest. The logistic part showed that absence of metastasis significantly increased the chance of node-negative breast cancer (P=0.027). The negative binomial part revealed an increase of 86% in the risk of a greater number of involved nodes in stage III breast cancer compared to stages I and II, suggesting that the patients were at a high risk (P=0.006).

Conclusion: Metastasis status and tumor grade significantly relate to the number of lymph nodes involved in breast cancer. Determining the factors associated with nodal involvement is crucial for the early diagnosis of breast cancer by clinicians.

Introduction

Breast cancer remains the most prevalent type of cancer among women worldwide (1-2). In Iran, breast cancer accounts for 24.4% of all cancers and contributes to 14.2% of cancer-associated mortality in women (3-4). The breast cancer incidence peaks in the fourth and fifth decades of life, i.e. about a decade earlier than the peak incidence of other countries (5-6). Despite the advances made in treatment strategies, breast cancer is a major global public health issue. Previous studies showed that the main causes for breast cancer could be aging, geographic influences, genetic predisposing, exposure to types of

[*] . Corresponding Author's Email: e.maraghi@gmail.com

radiation, early menstruation, number of pregnancies, first children after age 30, late menopause, hair coloring, stress and smoking (7-9). Besides identifying risk factors, cancer prevention should be take into account. Breast cancer is often accompanied by presence or absence of axillary lymph node involvement. Axillary lymph node metastasis has relation to risk of distant recurrence, staging and therapy (10, 11). Moreover, the number of involved nodes should also be clinically predicted to ascertain postoperative radiotherapy and chemotherapy (11-12).

The number of involved auxiliary lymph nodes as a discrete variable, are highly variable. Ordinary linear regression is not appropriate for modeling such discrete and highly variable outcomes. Recently, some studies explored a number of statistical models to examine model potential abilities for predicting the number of involved nodes (10 and 12). Poisson and negative binomial are commonly used count models in literature. The Poisson model assumes that the mean equals the variance (13). However, in many medical fields, count outcomes consist of excess zeroes that rule out this assumption (14-16). The negative binomial regression model, which assumes that the sample comprises two latent sub-groups, is employed to overcome this limitation (17). In 1992, Lambert introduced zero-inflated Poisson (ZIP) as a two-part model (18) that combined a latent binary distribution with the common Poisson model. In 1994, Greene proposed the ZINB model as an advance on ZIP (19). The ZINB model was found to be a more effective alternative to the ZIP model given the over dispersion and extra zeros in count outcomes (20).

As far as the authors' thorough review of literature suggests, the very few studies investigating the factors related to the number of nodes involved in breast cancer employed regression models for count outcomes (10 and 12). The present study sought to investigate the determinants of the number of involved lymph nodes in breast cancer using regression models. Given extra zeroes in the number of involved lymph nodes, zero-inflated count models were utilized for this purpose.

Materials and Methods

Participants

The present retrospective study recruited 165 eligible patients diagnosed with breast cancer during 2005-2015 referring to Shafa Hospital affiliated to Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. This study was approved by the Ethics Committee of the university (IR.AJUMS.REC.1398.402).

The number of involved lymph nodes was considered as the outcome and age, progesterone receptor status (positive/negative), status of human epidermal growth factor receptor2 (HER2) (positive/negative), tumor size, estrogen receptor status (positive/negative) and tumor grade (I+II/III) were taken as the main covariates. According to extra zeroes in the number of involved lymph nodes, zero-inflated count models were utilized for investigating the determinants of the number of lymph nodes involved in patients with breast cancer.

Statistical analyses

The continuous variables were expressed as mean values and the categorical data as frequency and relative frequency. The over-dispersion test and dispersion parameter were used to identify over-dispersion. The Poisson, negative binomial, ZIP and ZINB regression models were compared in terms of goodness-of-fit using the AIC and other goodness of fit indices such as log-likelihood, BIC and chi square goodness of fit. Statistical analyses were performed in Stata 16. A $P<0.05$ was considered as the level of statistical significance. Poisson regression, NB regression, ZIP model and ZINB model were useful for number of involved lymph nodes as a count outcome. Zero-inflated regression models interpreted as two-part models, consisting of both binary and count model sections in order to account for excess zero counts. The explored statistical models (i.e. P, NB, ZIP and ZINB) are defined as follows:

Poisson regression model

Let Yi denote the number of involved lymph nodes for the ith women. Assume that Yi follows a Poisson distribution with mean γi. So, the probability of observing any specific count Yi is:

$$f(\mathrm{y_i}|\mathrm{x_i}) = \frac{\mathrm{e}^{-\gamma_\mathrm{i}}\gamma_\mathrm{i}^{\mathrm{y_i}}}{\mathrm{y_i}!}$$

Suppose a linear relationship between covariates and mean:

$$E(y_i|\boldsymbol{x_i}) = \gamma_i = exp(\boldsymbol{x'_i\beta})$$

So, the log likelihood for Poisson model will be:

$$\ln L(\beta) = \sum_{i=1}^{n}\{y_i\boldsymbol{x'_i\beta} - exp(\boldsymbol{x'_i\beta}) - lny_i!\}$$

Equality of mean and variance is the assumption in Poisson regression. The Poisson regression model is not often well-suited for use in real datasets. Because count outcomes usually exhibit over-dispersion and/ or an excess number of zeroes. Negative Binomial regression model is the best alternative to Poisson model in the former issue.

Negative Binomial regression model

Let Yi be the number of involved lymph nodes for the ith women, having a Negative Binomial distribution with parameters α and μ (probability of having involved lymph nodes). Hence, the probability of observing any specific count of Yi is:

$$f(y|\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}$$

$$\alpha \geq 0, \qquad y = 0,1,2,\ldots$$

If linear relationship between covariates and mean established as $\mu = exp(\boldsymbol{x'_i\beta})$, then the log likelihood for Negative Binomial model will be

$$ln\left(\frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right) = \sum_{j=0}^{y-1} ln(j+\alpha^{-1})$$

$$lnL(\alpha,\beta) = \sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1} ln(j+\alpha^{-1})\right) - lny_i! - (y_i+\alpha^{-1})ln\big(1+\alpha exp(\boldsymbol{x'_i\beta})\big) + y_i ln\alpha + y_i\boldsymbol{x'_i\beta}\right\}$$

In some situations, excess zeroes is the cause of over-dispersion. Lambert suggested the use of zero inflated regression models as one of the alternatives to Poisson and Negative Binomial regression models (13).Whenever the excess zeroes are a combination of structural zeroes and sampling zeroes, the Zero-Inflated models are mostly applied. If we consider the zero involved lymph nodes in our studied population is a combination of structural zeroes (i.e. patients with a low risk of involved lymph nodes) and sampling zeroes (i.e. patients with a high risk of involved lymph node), then we can use zero-inflated models to capture the over-dispersion due to excess zeroes and unobserved heterogeneity among women with breast cancer.

In the zero-inflated models, structural zeroes are described through logistic regression. Sampling zeroes are described through the zero-inflated count model. In this study, sampling zeroes involved lymph nodes implies a woman who is at a "high-risk" of lymph nodes' involvement but does not have an involved lymph nodes due to chance.

Zero-Inflated Poisson model

The ZIP model refers to raw dataset as a mixture including an all-zero subset and a subset following Poisson distribution. The ZIP model supposes that

$$\Pr(\mathrm{Y=y}) = \begin{cases} \psi + (1-\psi)\mathrm{e}^{-\mu} & \mathrm{y}=0 \\ (1-\psi)\dfrac{\mathrm{e}^{-\mu}\mu^{\mathrm{y}}}{\mathrm{y}!} & \mathrm{y}=1,2,\ldots \end{cases}$$

$$\text{Logit }(\psi_\mathrm{i}) = \mathrm{Z'_i}\gamma$$

$$\text{Log}(\mu_\mathrm{i}) = \mathrm{X'_i}\lambda$$

$$L(\mu,\psi|y)=\prod_{i=1}^{n}\left[\left(\psi_i+(1-\psi_i)e^{-\mu_i}\right)I(y_i=0)\left((1-\psi_i)\frac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}\right)I(y_i\geq 0)\right]$$

Zero-Inflated Negative Binomial model

The ZINB model refers to raw dataset as a mixture including an all-zero subset and a subset following NB distribution. The probability density function of the ZINB model is:

$$P(Y=y)=\begin{cases}\psi+(1-\psi)\left(\frac{\delta}{1+\delta}\right)^{\mu} & y=0\\ (1-\psi)\frac{\Gamma(y+\mu)}{y!\,\Gamma(\mu)}\left(\frac{\delta}{1+\delta}\right)^{\mu}\left(\frac{1}{1+\delta}\right)^{y} & y=1,2,\ldots\end{cases}$$

$$\mathrm{Log}(\mu_i)=X_i'\lambda$$

$$\mathrm{Logit}\ (\psi_i)=Z_i'\gamma$$

$$L(\mu,\psi|y)=\prod_{i=1}^{n}\left(\psi_i+(1-\psi_i)\left(\frac{\delta}{1+\delta}\right)^{\mu_i}\right)I(y_i=0)$$

$$\times\left((1-\psi_i)\frac{(y_i+\mu_i)}{y_i!(\mu_i)}\left(\frac{\delta}{1+\delta}\right)^{\mu_i}\left(\frac{1}{1+\delta}\right)^{y_i}\right)\ I(y_i\geq 0)$$

Results

The final analyses were performed on 165 eligible women in 2005-2015. The patients were 27-75 years old and had a mean age of 46.40±9.94 years. The median count of involved lymph nodes was found to be 2.00 (IQR=0-6).

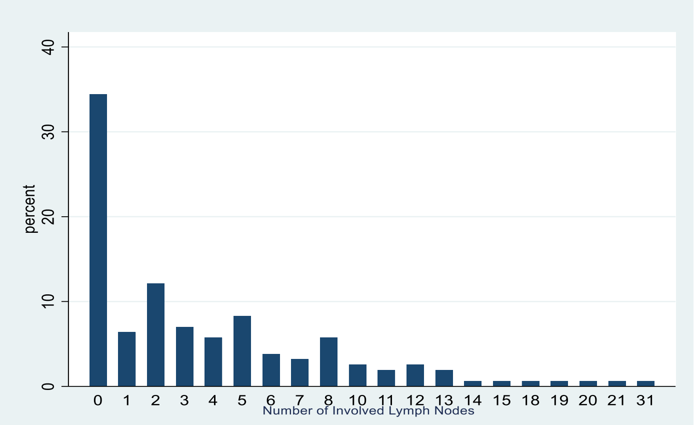

Figure1. Distribution of number of involved lymph nodes

Lymph nodes were involved in 63.64% (n=105) of the patients. Table 1 presents the details of the participants.

Table 1: Characteristics of the patients

Characteristics	Statistic
Age at diagnosis (year)	46.40±9.94**
Tumor Size (cm)	3.79±2.00**
Number of involved lymph nodes	2.00 (0-6)*
Tumor grade: III	89.0 (53.9)
Estrogen receptor: positive	116.0 (70.3)
Progesterone receptor: positive	106 (65.84)

HER2: positive	80.0 (48.5)

Given an alpha dispersion parameter of 0.525 [95% CI: 0.281-0.979], the number of involved lymph nodes was over-dispersed. According to table 2, observing the lowest AIC in the ZINB model suggested its highest goodness of fit.

Table 2: Goodness of fit of the regression models

	P	NB	ZIP	ZINB
AIC	990.77	706.26	808.16	698.99
Log-likelihood	-486.38	-343.1326	-386.08	-330.49
BIC	1017.44	735.89	861.49	755.28
Chi square goodness of fit	138.80	21.51	46.42	12.71

Table 3 estimates regression coefficients of different risk factors for both parts of the Poisson, Negative Binomial, ZIP and ZINB models. The logistic part showed significant increases in the chance of node-negative breast cancer caused by absence of metastasis (P=0.027). The negative binomial part revealed an increase of 86% in the risk of a greater number of involved nodes in stage III breast cancer compared to stages I and II, suggesting that the patients were at a high risk (P=0.006). The risk of developing a large number of positive nodes in the estrogen receptor-positive patients was 1.07 times that in the estrogen receptor-negative patients. The involvement of positive nodes in the HER2-positive patients was 1.23 times that in the HER2-negative patients.

Table 3: Regression models for count outcome's coefficients for the number of involved lymph nodes.

Poisson regression coefficients for the number of involved lymph nodes

Predictor	Risk Ratio (95% CI)	P-value
Age at diagnosis (year)	0.99 (0.98,1.003)	0.233
Grade: III	1.94 (1.53,2.46)	< 0.0001
Pathology: ductal	1.53 (1.07,2.18)	0.019
Metastasis: positive	1.63 (1.37,1.94)	<0.0001
Tumor size (cm)	1.08 (1.03,1.12)	<0.0001
HER2: positive	1.15 (0.96,1.38)	0.106
Progesterone receptor: positive	1.31 (0.95,1.81)	0.097
Estrogen receptor: positive	1.24 (0.87,1.75)	0.217

Negative Binomial regression coefficients for the number of involved lymph nodes

Predictor	Risk Ratio (95% CI)	P-value
Age at diagnosis (year)	1.0002 (0.97,1.02)	0.982
Grade: III	2.03 (1.24,3.34)	0.005
Pathology: ductal	1.56 (0.71,3.41)	0.265
Metastasis: positive	1.59 (0.99,2.54)	0.052
Tumor size (cm)	1.10 (0.97,1.25)	0.131
HER2: positive	1.14 (0.70,1.84)	0.583
Progesterone receptor: positive	1.36 (0.59,3.12)	0.460
Estrogen receptor: positive	1.28 (0.52,3.13)	0.581

Zero-Inflated Ppisson regression coefficients for the number of involved lymph nodes

Predictor	Logistic part Odds Ratio (95% CI)	P-value	Poisson part Risk Ratio (95% CI)	P-value

Age at diagnosis (year)	1.008 (0.96,1.04)	0.678	0.99 (0.99,1.006)	0.692
Grade: III	0.60 (0.24,1.48)	0.273	1.67 (1.30,2.15)	<.0001
Pathology: ductal	0.69 (0.15,3.09)	0.637	1.45 (1.006,2.09)	0.046
Metastasis: positive	0.19 (0.06,0.53)	0.002	1.16 (0.98,1.38)	0.080
Tumor size (cm)	0.79 (0.60,1.03)	0.093	1.02 (0.98,1.06)	0.200
HER2: positive	1.32 (0.55,3.11)	0.526	1.28 (1.06,1.53)	0.008
Progesterone receptor: positive	0.67 (0.13,3.39)	0.637	1.25 (0.92,1.71)	0.143
Estrogen receptor: positive	0.57 (0.10,3.03)	0.516	1.06 (0.76,1.47)	0.724

Zero-Inflated Negative Binomial regression coefficients for the number of involved lymph nodes

Predictor	Logistic part Odds Ratio (95% CI)	P-value	Negative-binomial part Risk Ratio (95% CI)	P-value
Age at diagnosis (year)	1.008 (0.95,1.06)	0.759	0.99 (0.97,1.01)	0.858
Grade: III	0.74 (0.22,2.45)	0.624	1.86 (1.19,2.90)	0.006
Pathology: ductal	0.62 (0.10,3.77)	0.612	1.41 (0.73,2.71)	0.296
Metastasis: positive	0.14 (0.03,0.80)	0.027	1.15 (0.79,1.67)	0.443
Tumor size (cm)	0.77 (0.55,1.08)	0.141	1.03 (0.94,1.13)	0.460
HER2: positive	1.29 (0.42,3.94)	0.654	1.23 (0.81,1.87)	0.312
Progesterone receptor: positive	0.56 (0.08,4.11)	0.571	1.23 (0.63,2.39)	0.539
Estrogen receptor: positive	0.59 (0.09,4.15)	0.604	1.07 (0.53,2.17)	0.841

Discussion

The present study used four popular models of count outcomes. Using zero-inflated models for analyzing outcomes with extra zeros results in unbiased and more efficient estimates in medicine. Given the common heterogeneity in medical research, effective and flexible models of zero-inflated types can assist physicians in making better decisions on patient-specific treatments. As a key prognostic and therapeutic factor in breast cancer (21), the number of involved lymph nodes should be predicted by physicians to improve health outcomes in patients with breast cancer. Given that this number as a count outcome often involves extra zeros that cause over-dispersion, zero-inflated models should be fitted to account for variability caused by extra zeros.

The ZINB model has the best fit in this study. This result is in accordance with that the distribution of the count outcome (number of involved nodes) is affected by over-dispersion due to excessive negative nodes and unobserved heterogeneity. So, worst fit of Poisson regression model is due to one source of over-dispersion.

This research pioneered the pattern analysis of nodal involvement in breast cancer by employing a dataset collected in Iran. The present findings showed involved nodes in 63.62% of the patients. Nodal involvement was also reported in 70.06% of a sample comprising Iranian patients (22). Moreover, nodal positivity rates were found higher in Indian patients (23). The present study found the HER2-positive status to be associated with increases in the risk of nodal involvement and the number of involved nodes. Moreover, the higher the grade, the higher the risk of increased number of involved nodes. These results are consistent with those reported in literature (24-26). The present study limitations comprised missing some important covariates data due to lack of a well-established medical registry, lack of similar works and being a secondary analysis using a dataset that was not specifically designed.

Conclusion

Metastasis status and tumor grade significantly relate to the involvement of lymph nodes in breast cancer. The early diagnosis of breast cancer requires that clinicians focus on the factors associated with nodal involvement.

Acknowledgments

The present study was supported by Ahvaz Jundishapur University of Medical Sciences (Project number: U-98112).

Conflicts of Interest

The authors declared no conflicts of interest as for the publication of the present paper.

References

1.Pelizzari G, Basile D, Zago S, Lisanti C, Bartoletti M, Bortot L, Vitale MG, Fanotto V, Barban S, Cinausero M, Bonotto M. Lactate Dehydrogenase (LDH) Response to First-Line Treatment Predicts Survival in Metastatic Breast Cancer: First Clues for a Cost-Effective and Dynamic Biomarker. Cancers. 2019 Sep;11(9):1243.

2.Presti D, Quaquarini E. The PI3K/AKT/mTOR and CDK4/6 Pathways in Endocrine Resistant HR+/HER2− Metastatic Breast Cancer: Biological Mechanisms and New Treatments. Cancers. 2019 Sep;11(9):1242.

3.Kolah DS, Sajadi A, Radmard AR, KHADEMI H. Five common cancers in Iran.

4.Abass MO, Gismalla MD, Alsheikh AA, Elhassan MM. Axillary lymph node dissection for breast cancer: efficacy and complication in developing countries. Journal of global oncology. 2018 Oct;4:1-8.

5. M. E. Akbari and G. Mohammadi, Women’s Cancers of Iran, Mohsen Publications, Tehran, 2014, (Farsi).

6. CDC Cancer Office, “National Cancer Registry Report 2007-8,” Tech. Rep., Tehran: Ministry of Health, Treatment and Education of Iran, Iran, 2009.

7.Hajian K, Gholizadehpasha A, Bozorgzadeh SH. Association of obesity and central obesity with breast cancer risk in pre-and postmenopausal women. Journal of Babol university of medical sciences. 2013 May 10;15(3):7-15.

8. Hajizadeh N. Incidence rate of breast cancer in iranian women, trend analysis from 2003 to 2009.

9.YektaKooshali MH, Esmaeilpour-Bandboni M, Sharami H, Alipour Z. Survival Rate and Average Age of the Patients with Breast Cancer in Iran: Systematic Review and Meta-Analysis. J Babol Univ Med Sci. 2016;18(8):29-40.

10. Swain PK, Grover G, Chakravorty S, Goel K, Singh V. Estimation of Number of Involved Lymph Nodes in Breast Cancer Patients using Bayesian Regression Approach.2017

11. Dwivedi AK, Dwivedi SN, Deo S, Shukla R, Kopras E. Statistical models for predicting number of involved nodes in breast cancer patients. Health. 2010 Jul;2(7):641.

12. Cui X, Wang N, Zhao Y, Chen S, Li S, Xu M, Chai R. Preoperative prediction of axillary lymph node metastasis in breast cancer using radiomics features of DCE-MRI. Scientific reports. 2019 Feb 19;9(1):1-8.

13. Cameron AC, Trivedi PK. Regression analysis of count data: Cambridge university press; 2013.

14. Mohammadi T, Kheiri S, Sedehi M. The application of zero-inflated count regression models for identifying main factors on the number of blood donor deferral in Shahrekord. J Shahrekord Univ Med Sci. 2016; 18(5): 26-35.

15. Bakhshi E, Yazdanipour MA, Rahgozar M, Ghorbani Z, Deghatipour M. Overall Effects of Risk Factors Associated with Dental Caries Indices Using the Marginalized Zero-Inflated Negative Binomial Model. Caries research. 2019 Jan 1;53(4):1-6.

16. Cantarero Prieto D, Pascual Sáez M, Lera Torres JI. Socioeconomic determinants and health care utilization among elderly people living in Europe:

Evidence from the Survey of Health, Ageing and Retirement. 2018.

17. Hilbe JM. Negative binomial regression: Cambridge University Press; 2011.

18. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1):1-14.

19. Greene WH. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. 1994.

20. Kamalja KK, Wagh YS. Estimation in zero-inflated Generalized Poisson distribution. Journal of Data Science. 2018 Jan 1;16(1):183-206.

21. Mohammed AA. Predictive factors affecting axillary lymph node involvement in patients with breast cancer in Duhok: Cross-sectional study. Annals of Medicine and Surgery. 2019 Aug 1;44:87-90.

22. Keihanian S, Koochaki N, Pouya M, Zakerihamidi M. Factors Affecting axillary lymph node involvement in patients with breast cancer. Tehran University Medical Journal TUMS Publications. 2019 Nov 10;77(8):484-90.

23. Chakraborty A, Bose CK, Basak J, Sen AN, Mishra R, Mukhopadhyay A. Determinants of lymph node status in women with breast cancer: A hospital based study from eastern India. The Indian journal of medical research. 2016 May;143(Suppl 1):S45.

24. Guern AS, Vinh-Hung V. Statistical distribution of involved axillary lymph nodes in breast cancer. Bulletin du cancer. 2008 Apr 1;95(4):449-55.

25. Kendal WS. Statistical kinematics of axillary nodal metastases in breast carcinoma. Clinical & experimental metastasis. 2005 Apr 1;22(2):177-83.

26. Schaapveld M, Otter R, de Vries EG, Fidler V, Grond JA, van der Graaf WT, de Vogel PL, Willemse PH. Variability in axillary lymph node dissection for breast cancer. Journal of surgical oncology. 2004 Jul 15;87(1):4-12.