Influence of Covariates on Involved Lymph Nodes in Primary Breast Cancer Patients: Mixture Distribution Zero-Inflated Modeling Methodological Framework

Madiha Liaqat[1], Shahid Kamal[2], Florian Fischer[3,4], Nadeem Zia[5]

[1] College of Statistical and Actuarial Sciences (CSAS), University of the Punjab, Lahore, Pakistan

[2] College of Statistical and Actuarial Sciences (CSAS), University of the Punjab, Lahore, Pakistan

[3] Institute of Public Health, Charité – Universitätsmedizin Berlin, Berlin, Germany

[4] Institute of Gerontological Health Services and Nursing Research, Ravensburg-Weingarten University of Applied Sciences, Weingarten, Germany

[5] Department of Oncology and Radiotherapy, Mayo Hospital, Lahore, Pakistan

Corresponding author:

Dr. Florian Fischer

Charité – Universitätsmedizin Berlin

Institute of Public Health

Charitéplatz 1

10117 Berlin

E-Mail: florian.fischer1@charite.de

Abstract

Background: Involvement of lymph nodes has been an integral part of breast cancer prognosis and survival. This study aimed to explore factors influencing on the number of auxiliary lymph nodes in women diagnosed with primary breast cancer by choosing an efficient model to assess excess of zeros and over-dispersion presented in the study population.

Methods: The study is based on a retrospective analysis of hospital records among 5,196 female breast cancer patients in Pakistan. Zero-inflated Poisson and zero-inflated negative binomial modeling techniques are used to assess the association between under-study factors and the number of involved lymph nodes in breast cancer patients.

Results: The most common breast cancer was invasive ductal carcinoma (54.5%). Patients median age was 48 years, from which women aged 46 years and above are the majority of the study population (64.8%). Examination of tumors revealed that over 2,662 (51.2%) women were ER-positive, 2,652 (51.0%) PR-positive, and 2,754 (53.0%) were Her2.neu-positive. The mean tumor size was 3.06 cm and histological grade 1 (n=2021, 38.9%) was most common in this sample.

The model performance was best in the zero-inflated negative binomial model. Findings indicate that most factors related to breast cancer have a significant impact on the number of involved lymph nodes. Age is not contributed to lymph node status. Women having a larger tumor size suffered from greater number of involved lymph nodes. Tumor grades 11 and 111 contributed to higher numbers of positive lymph node.

Conclusions: Zero-inflated models have successfully demonstrated the advantage of fitting count nodal data when both “at-harm” (lymph node involvement) and “not-at-harm” (no lymph node involvement) groups are important in predicting disease on set and disease progression. Our analysis showed that ZINB is the best model for predicting and describing the number of involved nodes in primary breast cancer, when overdispersion arises due to a large number of patients with no lymph node involvement. This is important for accurate prediction both for therapy and prognosis of breast cancer patients.

Keywords: Oncology, Mamma cancer, Nodal involvement, Count model, Zero-Inflation

Background

Breast cancer, a commonly diagnosed malignant cancer entity in females, represents a major public health issue worldwide [1]. Previous studies have shown large absolute numbers of incident breast cancer cases in developing countries, in which abnormal growth starts in breast tissues with the risk of spreading to other body parts [2]. This malignancy is classified into two major types, ductal and lobular carcinoma: Ductal carcinoma – which most breast cancers belong to – starts in the ducts; lobular carcinoma starts in the milk-producing parts of the breast (lobules). Significant prognostic factors of poor survival are higher age, nodal involvement, higher tumor grade, advanced clinical stage, greater tumor size, and metastasis [3].

Although new tumor markers have been identified and different tests are recommended, the condition is not good for lower economic countries [4]. Due to lack of funding, the tumor diagnosis and prognosis in public hospitals in developing countries is still limited to lymph node status, tumor size, and grade [4,5], also, the presence or absence of auxiliary lymph nodes has been recognized as an important predictor of breast cancer risk. Studies showed node-positive patients had lower survival rates than node-negative ones [6]. Furthermore, a higher number of positive lymph node involvement contributes to an increased risk of complications [7–9]. Many studies show the association between various factors and the progression of breast cancer; all of them highlight the importance of lymph node involvement in breast carcinoma [10,11].

For breast cancer prediction, modeling of non-negative whole integers is done through generalized linear modeling framework, by accounting non-normal response average linkage to the predictors. From the exponential family distributions, any distribution of projection of non-negative count response is used [12]. The number of involved lymph nodes falls under the category of count data, which cannot be generally approximated by a normal distribution. In the

context of count response variable, Poisson, negative binomial, Poisson log-normal and other mixtures of distributional models are available [13–15]. Among the negative binomial and Poisson regression models, binomial is mostly applied to avoid observed heterogeneity and clustering, while Poisson count model has the limitations of equal mean and variance in real-world scenarios [16,17].

In epidemiological investigations, excessive zeros are a special case of count data analysis, which cannot be applied through conventional count techniques. Zero-inflated distributional modeling has proved usefulness in this regard, by dividing the regime into zeros and non-zeros. These divided regions mostly called perfect and imperfect states, respectively [18]. A large variety of literature is available on zero-inflated response treated with zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB) modeling techniques [19–24].

Lymph node involvement in breast cancer is strongly associated with many factors. Although count models are suitable for modeling lymph node counts, but distributions of lymph node counts are characterized by a large number of zeroes, when there is no lymph node involvement at the initial diagnostic stage of breast cancer, that's why zero inflated modeling framework has replaced by count modeling to handle such "excess zeros". In studies where both excess zeros and over-dispersion occurs, negative binomial zero-inflated models are applied to get efficient estimates of parameters, rather than inflated Poisson [25]. Another common issue arises in understanding and projecting excessive zeros in studies so an appropriate model can be chosen to assess excessive zeros in the first place, by considering false negative rate [26].

This study aims to quantify the risk of lymph node involvement in women diagnosed with primary breast cancer, by applying descriptive statistics and other tests to check excessive zeros in the data set. Complete modeling methodology is presented and applied to study breast cancer

disease and develop better intervention strategies related to such data type. For doing so, we used retrospective data from hospital records of breast cancer patients in Pakistan.

Methods

Study design

This study is based on a retrospective analysis of data from 5,196 primary breast cancer women who registered at Mayo hospital Lahore, Pakistan, from 2013 to 2019. This data taken from hospital records include information about the age at diagnosis, cancer type, tumor size, histological grade, and molecular markers (ER, PR, Her2.neu). The number of involved lymph nodes was taken as the response variable. Complete information of predictors and response were available for all selected cases. Exclusion criteria were incomplete information, patients who had a secondary tumor or had metastasis from other organs to the breast at the time of registration, unknown pathological nodal status (Nx), immeasurable primary tumor (Tx), and Paget's disease of the nipple without tumor. The association between the understudy factors mentioned above and the number of involved nodes has been assessed using inflated count modeling.

Modeling framework

By representing count lymph nodes, the response variable $Y_j = 1,2,3,....,n$ has a probability mass function (PMF) of Poisson distribution with parameter θ,

$$f(y_j, \lambda_j) = \frac{e^{-\lambda_j} \lambda_j^{y_j}}{y_j!} \qquad y_j = 0,1,2,3,... \tag{1}$$

In this scenario, likelihood function takes the form

$$l(\lambda) = \ln(\lambda; y) = \sum_{j=1}^{n} \left\{ y_j \ln(\lambda_j) - \lambda_j - \ln(y_j!) \right\} \tag{2}$$

The function $\ln(\lambda_j) = \eta_j = x_j^T \alpha$ linked Y_j and matrix of explanatory variables X [12].

Poisson distribution has conditions of independent events and equality of conditional mean and variance [13].

If non-negative whole integer response follows negative binomial distribution the function is given by,

$$f(y_j;\lambda_j,\tau) = \frac{\Gamma\left(y_j + \frac{1}{\tau}\right)}{\Gamma\left(\frac{1}{\tau}\right) y_j!}\left(1+\tau\lambda_j\right)^{-\frac{1}{\tau}}\left(1+\frac{1}{\tau\lambda_j}\right)^{-y_j},\ y_j = 0,1,2,3,.... \quad (3)$$

Mean and variance are θ_j and $\theta_j(1+\tau\theta_j)$, respectively. Being a dispersion parameter $\tau \to 0$ negative binomial becomes a Poisson distribution [14].

Many fields of research involve excess zero counts Such data is known as zero-inflated data, in which the population under study is divided into two latent classes, with most observations being zero. Researchers have developed methods to analyze excessive zeroes data in various research disciplines such as agriculture, ecology, economy, health, tourism, transportation, manufacturing, and others. Zero-inflation has a two-stage modeling framework, where numbers are modeled through count distribution and observations at zero through binomial realization [25-28].

Lambert (1992) proposed the theory and application of zero-inflated Poisson (ZIP) modeling by a two-way process, first excess zeros and second zeros occur in Poisson distribution [19]. ZIP data has the general form,

$$Y_j \sim \begin{cases} 0 & probability \quad \pi_j \\ Poisson(\lambda_j) & probability \quad 1-\pi_j \end{cases} \quad (4)$$

Here, $Y_j, j = 1,2,3,....,n$, π_j is the probability of extra zeros while λ_j is the average of non-extra zeros subpopulation.

It yields two states mixture distribution with PMF,

$$\Pr(Y_j = y_j) = \begin{cases} \pi_j + (1-\pi_j)e^{-\lambda_j}, y_j = 0 \\ (1-\pi_j)\dfrac{e^{-\lambda_j}\lambda^{y_j}}{y_j!}, y_j > 0 \end{cases} \tag{5}$$

Parameters π_j and λ_j are defined as $\text{logit}(\pi_j) = Z_j\gamma$ and $\log(\lambda_j) = X_j\alpha$, where $\gamma = (\gamma_1, \gamma_2, \gamma_3, ..., \gamma_{p1})'$, $\alpha = (\alpha_1, \alpha_2, \alpha_3, ..., \alpha_{p2})'$ and Z_j, X_j are $(1\times p_1)$ and $(1\times p_2)$ vectors of predictors for the j^{th} unit. For estimation purpose, the likelihood for ZIP has the form,

$$l = l(\theta, \pi; y) = \sum_j \left\{ I_{(y_{j=0})}[\ln(\pi_j + (1-\pi_j)e^{-\lambda_j}] + I_{(y_{j>0})}[\ln(1-\pi_j) - \lambda_j + y_j \ln\lambda_j - \ln(y_j!)] \right\} \tag{6}$$

Where, $I(.)$ is the event indicator variable.

Inflated negative binomial [29] has the probability density function (PDF),

$$p(Y_j = y_j) = \begin{cases} \pi_j + (1-\pi_j)(1+\tau\lambda_j)^{-\frac{1}{\tau}}, y_j = 0 \\ (1-\pi_j)\dfrac{\Gamma\left(y_j + \dfrac{1}{\tau}\right)}{y_j!\Gamma\left(\dfrac{1}{\tau}\right)}\left(1+\tau\lambda_j\right)^{-\frac{1}{\tau}}\left(1+\dfrac{1}{\tau\lambda}\right)^{-y_j}, y_j > 0 \end{cases} \tag{7}$$

$$l = l(\tau, \lambda_j, \pi_j; y_j) = \sum_{j=1}^{n} \left\{ I_{(y_j=0)} \ln\left((\pi_j + (1-\pi_j)(1+\tau\lambda_j)^{-\frac{1}{\tau}} \right) + I_{(y_j>0)} \ln\left((1-\pi_j)\frac{\Gamma(y_j + \frac{1}{\tau})}{y_j!\Gamma(\frac{1}{\tau})}(1+\tau\lambda_j)^{-\frac{1}{\tau}}(1+\frac{1}{\tau\lambda_j})^{-y_j} \right) \right\}$$

(8)

As our data set has covariates, the ZIP likelihood has the form,

$$l = l\ln(y_j \mid x_j, z_j) = \sum_{j=1}^{n} \left\{ I_{(y_j=0)} \ln\left[(\exp z_j^T y) + \exp(-\exp(x_j^T\alpha)) \right] + I_{(y_j>0)} \left[y_j x_j^T\alpha - \exp(x_j^T\alpha) \right] - \ln(1+\exp(z_j^T y)) \right\}$$

134 (9)

135 Similarly, the ZINB likelihood with covariates is given by,

136 $l = l(y_j \mid x_j, z_j)$

137 $$= \sum_{j=1}^{n} \left\{ I_{(y_j=0)} \left[\ln\left(\exp(z_j^T y) + (1+\tau \exp(x_j^T \alpha)^{-\frac{1}{\tau}} \right) \right] + I_{(y_j>0)} \sum_{j=1}^{n} \left[j \sum_{k=1}^{y_j} \ln\left(\tau y_j - \tau k + 1\right) - \ln\left(1+\exp(\exp(z_j^T y))\right) - \ln(y_j!) - \left(y_j + \frac{1}{\tau} \right) \ln\left(1+\tau \exp(x_j^T \alpha)\right) + y_j x_j^T \alpha \right] \right\}$$

138 (10)

139 Estimates are obtained by taking first and second derivatives with respect to unknown
140 parameters.

141 We applied the approach to model the effect of different factors related to breast cancer on the
142 number of involved lymph nodes. The covariates found to be significant in univariate analysis
143 with any of the regressions were included in all the regression models to maintain the
144 comparative findings. We first applied Poisson and negative binomial count models and then ZIP
145 and ZINB models. The Vuong test was performed to check inflated models' improvement over
146 standard parametric count models [28,30]. The general statistical model has the form,

147 $Count(LymphNodes) = \alpha_0 + \alpha_1 TumorType + \alpha_2 Age + \alpha_3 TumorGrade + \alpha_4 ER +$

148 $\alpha_5 PR + \alpha_6 Her2.neu + \alpha_7 TumorSize$

149 (11)

150 Where count is the response variable (number of involved lymph nodes), which can follow many
151 distributions, like Poisson, Quasi-Poisson, negative binomial and others. Here we applied two
152 distributional count models: Poisson and negative binomial. For the Vuong test [31], we assumed
153 Poisson and negative binomial regression models, while for the binary component of zero-
154 inflated $2^7 = 128$ different choices occurred.

Zero inflated models reduced bias in our estimates by accounting non-normality due to occurrence of large number of zeroes in our data set. Also, Akaike information criterion (AIC) was criteria used to check the performance of zero-inflated models [32]:

$$AIC = -2\ln L + 2p \tag{12}$$

Where L denotes the fitted log likelihood and p the number of parameters, AIC is highly recommended model selection criteria as it takes into account both goodness of fit and complexity of the model.

For all analyses, the level of significance was chosen at 5%. R 3.6.2 package was used for data analysis [33].

Results

Sample characteristics and lymph node involvement

Table 1 shows the characteristics of female breast cancer patients, which may act as predictive factors on the number of lymph nodes. The tumor is divided into invasive ductal carcinoma (n=2,832; 54.5%) and other types (n=2,364; 45.5%). The median age at diagnosis was 48 and it was divided into three categories ≤35 years (n=736, 14.2%), 36–45 years (n=1,092; 21.0%), and ≥46 years (n=3368; 64.8%). The patients were almost equally distributed among the histological grades. About half of the patients were ER positive (51.2%), PR positive (51.0%), and Her2.neu positive (53.0%). The majority of the patients (70.2%) had a tumor size between 2 and 4.9 cm.

- Table 1 -

The plot of nodal frequency at each count is shown in Figure 1. Overall, 2,406 breast cancer patients (46,3%) had no lymph node involved, suggested inflated zero models. Among the 2,790 patients with positive nodes, factors are distributed among three counting categories 1–2 (n=656), 3–4 (n=1324), and ≥5 (n=810) of involved nodes (Table 2).

- Figure 1 -

- Table 2 -

Zero-inflation modeling

Zero inflated Poisson (ZIP) and negative binomial (ZINB) demonstrated significant ($p<0.05$) association of tumor grade, ER, PR and tumor size with counts of involved nodes, while tumor type, tumor grade, ER, PR, Her2.neu and tumor size have influence on zero excessive counts of no involvement of nodes (Table 3), because of not having prior information about predictors which might have strong association with positive nodes and/or excessive zeroes, we take same covariates on both components of two distributions mixture inflated models. Results showed estimated coefficients, standard errors and associated p-values for count and zero parts for ZIP and ZINB.

- Table 3 -

Vuong test [31], as a suitable suggested approach for both nested and non-nested models, favors between Poisson count and ZIP the latter (Z-statistic=-26.917, $p<0.001$), and between negative binomial and ZINB also the latter one (Z-statistic=-22.286, $p<0.001$). The model performance

based on AIC indicated that among linear (AIC=23710.43), counts (Poisson (AIC=20777.19); Negative Binomial (AIC =18773.01)) and zero inflated (ZIP (AIC=16658.36); ZINB (AIC=16559.15)) the ZINB model provides the best fit to the data, as reference of lower AIC value [32] (Table 4).

-Table 4 -

Due to the better fit of the ZINB model, these modeling results on factors significantly associated with nodal involvement are presented (Table 5). According to this, patients with a higher tumor grade have a larger number of positive lymph nodes, although strongly positive associated for grade 111 (OR=1.323, 95% CI: 1.248–1.402) and 11 (OR=1.002, 95% CI: 0.944─1.064) compared to grade 1. Patients being ER-negative (OR=0.952, 95% CI: 0.913–0.993) and PR-negative (OR=0.897, 95% CI: 0.856–0.939) have a lower likelihood for involved lymph nodes. Larger tumor sizes have positive association with a greater number of positive lymph nodes involved at primary diagnosed breast cancer in Pakistani women. When considering zero counts, it is also shown in Table 5 that women diagnosed with another tumor type than invasive ductal carcinoma are prone to have no involvement of lymph nodes (OR=6.869, 95% CI: 5.632–8.376). Patients with tumor grade 1 have positive correlation with not having lymph nodes than grades 11 (OR=0.375, 95% CI: 0.309–0.454) and 111 (OR=0.453, 95% CI: 0.372–0.550). ER-negative (OR=0.543, 95% CI: 0.460–0.642), PR-negative (OR=0.277, 95% CI: 0.229–0.334) and Her2.neu-negative (OR=0.429, 95% CI: 0.363–0.508) patients have greater probability of no nodal involvement tumor. Tumor size 2–4.9 (OR=0.497, 95% CI: 0.394–0.626) have zero counts

in respect of involved nodes than reference size ≤1.9, and bigger tumor size ≥5 (OR=0.036, 95% CI: 0.022–0.058).

-Table 5 -

Discussion

In our data, almost half of the study population with breast cancer showed negative nodal status, which means no involvement of lymph nodes at primary diagnosis of breast cancer. However, this can be due to an early stage of breast cancer which may develop to be more harmful after time passes. Furthermore, there is a probability of human error during data entry. For that reason, choosing the best fit is important in case of accounting false negative. The study population is divided into “at-harm” and “not-at-harm” groups in the context of nodal status, as patients with positive number of involved nodes are considered more critical than those who have no involvement of lymph nodes.

We have fitted standard Poisson and negative binomial regression models by incorporating time independent covariates. When it comes to describe the variability of counts, the Poisson model does not fit well, but negative binomial distribution serves a good fit for modeling counts with variability different from mean. This is depicted through the analysis, which resulted into higher AIC values due to ignoring the higher frequency of occurrences in zeroes. Whereas the estimates are correct, standard errors are wrong and unaccounted by the model. Therefore, the results are not displayed in this article.

Furthermore, lymph nodes negative status was in greater number than positive ones, which showa a strong evidence of applying zero inflated modeling techniques to get in depth study of

patients' health by analyzing data. Zero inflated models separately analyzed both groups count and excessive zeroes by mixture modeling technique.

ZIP is suggested to account excess zeros, but comes with the limitation of equal mean and variance for the count component [34]. Also for heavy tail with extra zero data ZINB model is recommended [29]. In summary, zero-inflated modeling is a better choice as compared to standard regression count models, for the type of study population we have analyzed. Therefore, it is highly recommended that researchers should fit standard and inflated models to choose the best fit for their dataset.

In terms of model selection criteria, score and likelihood ratio tests are suitable for comparing nested models [27]. Assuming a well fitted model, the ZINB model is likely to have provided an accurate representation for understudy dataset by AIC non-nested model selection method, a good way to provide tradeoff between complexity of applied model and good fit for study data.

Some limitations must be noted. First, the number of covariates added is dependent on the availability of data. Further important factors related to patients not included in this analysis may also affect the nodal status. Second, we were not able to account a longitudinal assessment that may reveal other aspects related to "at-harm" and "not-at-harm" populations. Third, we did not account for different binary components of zero-inflation due to a large number of available covariates. This choice of models needs to be empirically investigated in studies related to breast cancer in the future. Forth, simulation study can be conducted to strengthen our conclusions.

Apart from these future tasks, this study is trying to fill the statistical modeling gap to analyze patterns of nodal involvement in primary breast cancer patients, using a large data set collected in Pakistan. Mayo hospital Lahore is one of the best governmental hospital where patients come from all over Pakistan. We believe that our study successfully quantified the "at-harm"

population group by incorporating time independent covariates which are associated with the presence of involved lymph nodes. Zero-inflated models have successfully demonstrated the advantage of fitting count nodal data when both "at-harm" and "not-at-harm" groups are important in predicting disease on set and disease progression.

Conclusions

Accurate prediction of involved lymph nodes is a need for clinicians, as it is one of the most important therapeutic and prognostic factors to improve health outcomes of breast cancer patients. Our analysis showed that ZINB is the best model for predicting and describing the number of involved nodes in primary breast cancer, when overdispersion arises due to a large number of patients with no lymph node involvement. Furthermore, advanced tumor grade and larger tumor pose risks, which suggest urgent treatment.

Abbreviations

AIC	Akaike information criterion
CI	Confidence interval
ER	Estrogen receptor
Her2.neu	Human Epidermal Growth Factor Receptor 2
OR	Odds ratio
PDF	Probability density function
PMF	Probability mass function
PR	Progesterone receptor
p-value	Probability-value
SE	Standard error
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

DECLARATIONS

Ethics approval and consent to participate

No ethical approval needed, because the study is based on a retrospective analysis of hospital record data.

Consent for publication

Not applicable

Availability of data and materials

Data is available upon reasonable request from the corresponding author.

Competing interests

The authors declare that no competing interests exist.

Funding

This study received no funding.

Authors' contributions

The study was conceptualized by ML, supported by SK. NZ has been responsible for data acquisition. ML analyzed the data; SK and FF supervised this process. ML drafted the manuscript, SK and FF revised it critically for important intellectual content. All authors approved the final version of the manuscript.

Acknowledgements

We thank the Staff of Oncology and Radiology Department, Mayo Hospital, Lahore, who supported in data collection. We also wish to appreciate Dr. Abbas Khokar (Mbbs, FCPS), Head of Oncology Department from Mayo Hospital, Lahore, Pakistan, for all the efforts to organize patients' records so systematically.

We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

Authors' information

Not applicable

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424.
2. Fan L, Goss PE, Strasser-Weippl K. Current status and future projections of breast cancer in Asia. Breast Care. 2015;10:372–8.
3. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer. 2019;11:151–64.
4. Liaqat M, Shahid K, Fischer F, Fazil W. Performance Evaluation of Distributional Models to Analyze Random Right Censored Breast Cancer Failure Time Data. BMC Medical Research Methodology [under review].
5. Martin FT, O'Fearraigh C, Hanley C, Curran C, Sweeney KJ. The prognostic significance of nodal ratio on breast cancer recurrence and its potential for incorporation in a new prognostic index. Breast J. 2013;19:388–93.
6. Hong R, Dai Z, Zhu W, Xu B. Association between lymph node ratio and disease specific survival in breast cancer patients with one or two positive lymph nodes stratified by different local treatment modalities. PLoS One. 2015;29:e0138908.
7. Fisher B, Bauer M, Wickerham DL, Redmond CK, Fisher ER, Cruz AB, Foster R, Gardner B, Lerner H, Margolese R, Poisson R, Shibata H, Volk H. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. Cancer. 1983;52(9):1551–7.
8. Nottegar A, Veronese N, Senthil M, Roumen RM, Stubbs B, Choi AH, Verheuvel NC, Solmi M, Pea A, Capelli P, Fassan M, Sergi G, Manzato E, Maruzzo M, Bagante F, Koc M, Eryilmaz MA, Bria E, Carbognin L, Bonetti F, Barbareschi M, Luchini C. Extra-nodal extension of sentinel lymph node metastasis is a marker of poor prognosis in breast cancer patients: A systematic review and an exploratory meta-analysis. Eur J Surg Oncol. 2016;42:919–25.

9. Rao R, Euhus D, Mayo HG, Balch C. Axillary node interventions in breast cancer: a systematic review. JAMA. 2013;310:1385–94.
10. Ravdin PM, De Laurentiis M, Vendely T, Clark GM. Prediction of axillary lymph node status in breast cancer patients by use of prognostic indicators. Journal of National Cancer Institute. 1994;86(23):1771–5.
11. Olivotto IA, Jackson JSH, Mates D, Andersen S, Davidson W, Bryce CJ, Ragaz J. Prediction of axillary lymph node involvement of women with invasive breast carcinoma a multivariate analysis. Cancer. 1998;83(5):948–55.
12. Tang W, He H, Tu XM. Applied Categorical and Count Data Analysis. 2012. FL: Chapman & Hall/CRC.
13. Consul PC, Famoye F. Generalized Poisson regression model. Communications in Statistics (Theory & Method). 1992;2(1):89–109.
14. Hilbe J. Negative Binomial Regression. Cambridge, UK: Cambridge University Press. 2007.
15. Joe H, Zhu R. Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biometrical Journal. 2005;47:219–29.
16. Cox D. Some remarks on overdispersion. Biometrics. 1983;10:269–74.
17. Dean C. Testing for overdispersion in poisson and binomial regression models. Journal of the American Statistical Association. 1992;87:451–7.
18. Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med. 2002;21:1461–9.
19. Lambert D. Zero-inflated poisson regression with an application to defects in manufacturing. Technometrics. 1992;34:1–14.
20. Yau KK, Lee AH. Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. Stat Med. 2001;20:2907–20.
21. Agarwal DK, Gelfand A, Citron-Pousty S. Zero-inflated model with application to spatial count data. Environmental and Ecological Statistics. 2002;9:341–55.
22. Hur K, Hedeker D, Henderson W, Khuri S, Daley J. Modeling clustered count data with excess zeros in health care outcomes research. Health Services and Outcomes Research Methodology. 2002;3:5–20.

23. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J Biopharm Stat. 2006;16(4):463–81.
24. Buu A, Johnson N, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. Stat Med. 2011;30(18):2326–40.
25. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 2000;56(4):1030–9.
26. Tang W, Lu N, Chen T, Wang W, Gunzler DD, Han Y, Tu XM. On performance of parametric and distribution-free models for zero-inflated and over-dispersed count responses. Stat Med. 2015;34(24):3235–45.
27. Ridout MS, Hinde JP, Demetrio CGB. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics. 2001;57:219–23.
28. Atkins D, Gallop R. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. Journal of Family Psychology. 2007;21(4):726–35.
29. Yau KK, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biom J. 2003;45(4):437–52.
30. Gilthorpe MS, Frydenberg M, Cheng Y, Baelum V. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. Stat Med. 2009;28:3539–53.
31. Vuong, Quang H. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989;57(2):307–33.
32. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19:716–23.
33. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/. 2011.
34. Yang Z, Hardin JW, Addy CL. Testing overdispersion in the zero-inflated Poisson model. Journal of Statistical Planning and Inference. 2009;139:3340–53.

Figure 1: Nodal frequency count

Tables

Table 1: Characteristics of 5,196 patients with primary breast cancer

	n (%)
Tumor type	
IDC	2,832 (54.5%)
Other	2,364 (45.5%)
Age (in years)	
≤35	736 (14.2%)
36–45	1,092 (21.0%)
≥46	3,368 (64.8%)
Tumor grade	
1	2,021 (38.9%)
11	1,618 (31.1%)
111	1,557 (30.0%)
Estrogen receptor	
Positive	2,662 (51.2%)
Negative	2,534 (48.8%)
Progesterone receptor	
Positive	2,652 (51.0%)
Negative	2,544 (49.0%)
Her2.neu receptor	
Positive	2,754 (53.0%)
Negative	2,442 (47.0%)
Tumor size (in cm)	
≤1.9	963 (18.5%)
2–4.9	3,650 (70.2%)
≥5	583 (11.2%)

427 ***Table 2:*** Distribution of covariates with respect to positive number of involved nodes

Positive nodes (n)	**1–2 n (%)**	**3–4 n (%)**	**≥5 n (%)**	**Total n (%)**
Tumor type				
IDC	412 (20.1)	1,010 (49.3)	627 (30.6)	2,049 (100)
Other	244 (32.9)	314 (42.4)	183 (24.7)	741 (100)
Age (in years)				
≤35	84 (22.2)	181 (47.8)	114 (30.0)	379 (100)
36–45	123 (20.9)	289 (49.1)	177 (30.1)	589 (100)
≥46	449 (24.6)	854 (46.9)	519 (28.5)	1,822 (100)
Tumor grade				
1	214 (28.2)	458 (60.3)	88 (11.6)	760 (100)
11	251 (25.0)	518 (51.6)	235 (23.4)	1,004 (100)
111	191 (18.6)	348 (33.9)	487 (47.5)	1,026 (100)
Estrogen receptor				
Positive	305 (21.7)	682 (48.4)	421 (29.9)	1,408 (100)
Negative	351 (25.4)	642 (46.5)	389 (28.1)	1,382 (100)
Progesterone receptor				
Positive	247 (19.6)	530 (42.1)	481 (38.2)	1,258 (100)
Negative	409 (26.7)	794 (51.8)	329 (21.5)	1,532 (100)
Her2.neu receptor				
Positive	242 (21.6)	450 (40.2)	427 (38.2)	1,119 (100)
Negative	414 (24.8)	874 (52.3)	383 (22.9)	1,671 (100)
Tumor size (in cm)				
≤1.9	183 (73.5)	34 (13.7)	32 (12.9)	249 (100)
2–4.9	452 (22.8)	1,278 (64.4)	253 (12.8)	1,983 (100)
≥5	21 (3.8)	12 (2.2)	525 (94.1)	558 (100)

428

429 ***Table 3:*** Multivariate models to study the effect of time independent covariates on number of involved nodes in
430 primary breast cancer patients

Parameter	Zero-Inflated Poisson (ZIP) Estimate (SE) [p-value]	Zero-Inflated Negative Binomial (ZINB) Estimate (SE) [p-value]
Poisson/NB with log link		
Intercept	0.512 (0.065) [5.39e-15]	0.486 (0.071) [5.51e-12]
Tumor type		
IDC	Ref.	Ref.
Other	-0.004 (0.025) [0.870]	0.010 (0.028) [0.730]
Age (in years)		
≤35	Ref.	Ref.
36–45	-0.006 (0.032) [0.861]	0.005 (0.037) [0.897]
≥46	0.019 (0.027) [0.487]	0.019 (0.032) [0.540]
Tumor grade		
1	Ref.	Ref.
11	0.006 (0.027) [0.815]	0.002 (0.030) [0.940]
111	0.273 (0.026) [<2e-16]	0.280 (0.030) [<2e-16]
Estrogen receptor		
Positive	Ref.	Ref.
Negative	-0.047 (0.019) [0.011]	-0.049 (0.021) [0.021]
Progesterone receptor		
Positive	Ref.	Ref.
Negative	-0.087 (0.020) [2.07e-05]	-0.109 (0.024) [3.91e-6]
Her2.neu receptor		
Positive	Ref.	Ref.
Negative	-0.027 (0.019) [0.170]	-0.030 (0.023) [0.178]
Tumor size (in cm)		
≤1.9	Ref.	Ref.
2–4.9	0.705 (0.058) [<2e-16]	0.726 (0.061) [<2e-16]
≥5	1.622 (0.059) [< 2e-16]	1.654 (0.063) [<2e-16]
Log(theta)		2.908 (0.134) [<2e-16]
Binomial with logit link		
Intercept	1.300 (0.165) [3.68e-15]	1.263 (0.172) [1.80e-13]
Tumor type		
IDC	Ref.	Ref.
Other	1.845 (0.094) [<2e-16]	1.927 (0.101) [<2e-16]
Age (in years)		

Parameter	Zero-Inflated Poisson (ZIP) Estimate (SE) [p-value]	Zero-Inflated Negative Binomial (ZINB) Estimate (SE) [p-value]
≤35	Ref.	Ref.
36–45	0.073 (0.130) [0.573]	0.086 (0.134) [0.523]
≥46	0.027 (0.111) [0.805]	0.039 (0.115) [0.732]
Tumor grade		
1	Ref.	Ref.
11	-0.932 (0.093) [<2e-16]	-0.982 (0.080) [<2e-16]
111	-0.767 (0.096) [1.34e-15]	-0.792 (0.100) [1.82e-15]
Estrogen receptor		
Positive	Ref.	Ref.
Negative	-0.558 (0.081) [4.31e-12]	-0.610 (0.085) [7.48e-13]
Progesterone receptor		
Positive	Ref.	Ref.
Negative	-1.210 (0.089) [<2e-16]	-1.285 (0.096) [<2e-16]
Her2.neu receptor		
Positive	Ref.	Ref.
Negative	-0.828 (0.082) [<2e-16]	-0.846 (0.086) [<2e-16]
Tumor size (in cm)		
≤1.9	Ref.	Ref.
2–4.9	-0.728 (0.114) [1.74e-10]	-0.700 (0.118) [3.39e-09]
≥5	-3.362 (0.240) [<2e-16]	3.320 (0.246) [<2e-16]

433 ***Table 4:*** Comparison of fitted models by Akaike information criterion (AIC)

Model	**AIC**
Linear	23710.43
Poisson	20777.19
Negative Binomial	18773.01
Zero-Inflated Poisson	16658.36
Zero-Inflated Negative Binomial	16559.15

434

435 ***Table 5:*** Zero-Inflated Negative Binomial model

Parameter	OR	95% CI
Poisson/NB with log link		
Intercept	1.626	1.416–1.867
Tumor type		
IDC	1	
Other	1.010	0.955–1.067
Age (in years)		
≤35	1	
36–45	1.005	0.926–1.070
≥46	1.019	0.958–1.085
Tumor grade		
1	1	
11	1.002	0.944–1.064
111	1.323	1.248–1.402
Estrogen receptor		
Positive	1	
Negative	0.952	0.913–0.993
Progesterone receptor		
Positive	1	
Negative	0.897	0.865–0.939
Her2.neu receptor		
Positive	1	
Negative	0.970	0.928–1.014
Tumor size (in cm)		
≤1.9	1	
2–4.9	2.067	1.836–2.329
≥5	5.228	4.625–5.913
Binomial with logit link		
Intercept	3.536	2.527–4.952
Tumor type		
IDC	1	
Other	6.869	5.632–8.376
Age (in years)		
≤35	1	
36–45	1.090	0.838–1.417

Parameter	OR	95% CI
≥46	1.040	0.831–1.302
Tumor grade		
1	1	
11	0.375	0.309–0.454
111	0.453	0.372–0.550
Estrogen receptor		
Positive	1	
Negative	0.543	0.460–0.642
Progesterone receptor		
Positive	1	
Negative	0.277	0.229–0.334
Her2.neu receptor		
Positive	1	
Negative	0.429	0.363–0.508
Tumor size (in cm)		
≤1.9	1	
2–4.9	0.497	0.394–0.626
≥5	0.036	0.022–0.058

436