Journal of King Saud University – Science 34 (2022) 101932

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

Zero-inflated and hurdle models with an application to the number of involved axillary lymph nodes in primary breast cancer

Madiha Liaqat [a], Shahid Kamal [a], Florian Fischer [b,c,*], Nadeem Zia [d]

[a] College of Statistical and Actuarial Sciences (CSAS), University of the Punjab, Lahore, Pakistan
[b] Institute of Public Health, Charité – Universitätsmedizin Berlin, Berlin, Germany
[c] Institute of Gerontological Health Services and Nursing Research, Ravensburg-Weingarten University of Applied Sciences, Weingarten, Germany
[d] Department of Oncology and Radiotherapy, Mayo Hospital, Lahore, Pakistan

ARTICLE INFO

Article history:
Received 6 March 2021
Revised 15 February 2022
Accepted 25 February 2022
Available online 5 March 2022

Keywords:
Count data
Zero-inflated
Hurdle models
Information measures
Breast cancer
Nodal status

ABSTRACT

Objectives: This study aims to explore factors influencing the number of axillary lymph nodes in women diagnosed with primary breast cancer by choosing an efficient model to assess excess of zeros and overdispersion presented in the study population.
Methods: It is based on a retrospective analysis of hospital records among 5196 female breast cancer patients in Pakistan. Zero-inflated and hurdle modelling techniques are used to assess the association between under-study factors and the number of involved lymph nodes in breast cancer patients. Count data models including Poisson and negative binomial, zero-inflated models (zero-inflated Poisson and zero-inflated negative binomial), and hurdle models (hurdle Poisson and hurdle negative binomial) were applied. Performance evaluation of models was compared based on AIC, BIC, and zero counts capturing.
Results: The zero-inflated negative binomial model provided an acceptable fit. Findings indicate women who had a larger tumor in size suffered from the greater number of axillary involved lymph nodes from high-risk patients' group, also tumor grades II and III contributed to higher numbers of lymph nodes. Women's ages do not have any significant influence on nodal status.
Conclusions: Our analysis showed that the zero-inflated negative binomial 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.

1. Introduction

Count data usually occur in all disciplines, one approach to model such data is logistic regression after converting count into binary values. Such dichotomization conversion approach is suffered from the loss of information (Suissa and Blais, 1995). As a result, Poisson becomes the most adaptive regression model for analyzing count response data (Consul and Famoye, 1992), without dichotomization. The major drawback of Poisson distribution is the limitation of equal mean and variance, which cannot be fulfilled in many real-world scenarios. If neglected the assumption of equal mean and variance, Poisson regression produces biased estimates and misleading results (Winkelmann and Zimmermann, 1995), researchers have recommended the application of negative binomial distribution to relax this constraint; negative binomial distribution accounts for over-dispersion in count data by an additional

Abbreviations: AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; ER, Estrogen receptor; Her2, Human epidermal growth factor receptor 2; HNB, Hurdle negative binomial; HP, Hurdle Poisson; NB, Negative binomial; PR, Progesterone receptor; ZINB, Zero-inflated negative binomial; ZIP, Zero-inflated Poisson.

* Corresponding author at: Charité – Universitätsmedizin Berlin, Institute of Public Health, Charitéplatz 1, 10117 Berlin, Germany.
E-mail address: florian.fischer1@charite.de (F. Fischer).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2022.101932
1018-3647/

parameter (Hilbe, 2001). When overdispersion occurs due to a large number of zeros, analyzing such data using conventional count models (Poisson and negative binomial) is inappropriate. Zero-inflated models have proven their usefulness in this regard, by modelling the count response variable as a mixture of direct mass at the excess of zeros and count components. The zero-inflated Poisson (ZIP) model (Lambert, 1992; Böhning et al., 1999; Hall, 2000; Lee et al., 2001) is the most applied one in the literature of excess zeros count data. It assumes that the count component is displayed by the Poisson distribution. If count response data exhibit high variability due to excess of zeros and overdispersion, a negative binomial distribution is assumed to fit such data under mixture modelling technique, usually known as zero-inflated negative binomial (ZINB) model (Hall, 2000; Yesilova et al., 2010; Yau et al., 2003). One can also apply hurdle count models, if excess zeros only occur due to sampling variability in the data (Mullahy, 1986), it means hurdle count models consider the source of overdispersion only due to excess of zeros. The hurdle models, originally introduced by Mullahy (1986), are two-component models: the first component is modelled the probability of excess zeros and, the second component accounts for the non-excess zeros and non-zero counts. For the hurdle Poisson (HP) model, it is postulated that the positive count component is modelled via truncated Poisson distribution (Zorn, 1996; Moloas and Lesaffre, 2010). In case of overdispersion and excess zeros; the positive count component is modelled by the truncated negative binomial distribution, which is called the hurdle negative binomial (HNB) model (Rose et al., 2006).

Zero-inflated and hurdle along with other count models have been successfully employed in medical and health researches (Yau et al., 2003; Rose et al., 2006; Gilthorpe et al., 2009; Lee et al., 2006). The number of involved lymph nodes outcome variable falls under the category of count data, such count data exhibit many zero observations when there is no lymph node involvement at the initial diagnosis stage of breast cancer, which has a strong indication to apply zero-inflated and hurdle models. A study described patients may have a large number of negative nodal status at an early stage due to reporting error (Afifi et al., 2007). Furthermore, chances of false-negative recorded nodes cannot be neglected because of the non-dissection of complete axillary lymph nodes (Hur et al., 2002).

The main objective of the research reported in this article is to apply Poisson (P), negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), hurdle Poisson (HP), and hurdle negative binomial (HNB) models to analyze factors that may influence the number of involved nodes particularly in a case where there are chances of a high proportion of no involvement of lymph nodes exist. Poisson, negative binomial, and ZI and hurdle parameterizations for the Poisson and negative binomial distributions were fitted to breast cancer data. The complete modelling methodology is presented and results were compared.

2. Materials and methods

2.1. Study design

This study is based on a retrospective analysis of data from hospital records. Overall, 5196 primary breast cancer women who registered at Mayo hospital Lahore, Pakistan, from 2013 to 2019 are included in the analysis. Information about the age at diagnosis, cancer type, histological grade, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (Her2), and tumor size are included. The number of involved lymph nodes is taken as the response variable, negative nodal status data indicated by zeros. 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 assessed using zero-inflated and zero-hurdle models. Age at diagnosis, cancer type, tumor size, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (Her2), tumor grade, all predictors and their forms were chosen with the help of clinicians and oncologists.

In this data, age at diagnosis (in years) is divided into three categories ($\leq$35, 36—45, and $\geq$ 46). Age is mostly categorized in the literature related to breast cancer, because breast cancer risk factors have different effects on younger and older women. Different authors have defined a variety of age cutoffs due to a variety of reasons (Chollet-Hinton et al., 2016). Cancer type is represented by binary variable (0 = Lobular Carcinoma and other, 1 = Ductal Carcinoma), estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her2) are also represented by binary variables (0 = Negative, 1 = Positive), tumor grade is represented categorically (I, II, and III), and tumor size was classified into three categories ($\leq$1.9 cm, 2—4.9 cm, and $\geq$ 5 cm). Along with the aim of this study, which is the comparison of different count models, it is also of major interest to find out predictors that have a significant impact on the number of involved nodes.

2.2. Modelling framework

The modelling framework of count models are emerged from generalized regression models (Agresti, 1996), while generalized linear models are extended form of the simple linear regression model, which is written as:.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2, \cdots, \alpha_j x_j + \varepsilon \tag{1}$$

where, $x_1, x_2, \cdots, x_j$ are independent variables, α_0 is intercept and, $\alpha_1, \alpha_2, \cdots, \alpha_j$ are slope parameters, and, ε is a random error, which follows the normal distribution. This (1) can also be written as:

$$\lambda_j = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2, \cdots, \alpha_j x_j \tag{2}$$

The function λ_j is a linear function of the regressors, it can be denoted by $g\left(\mu_j\right)$, called the link function, which transforms the expectation of the response variable, and can also be written in log link function as:.

$$g\left(\mu_j\right) = log(\lambda_j) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2, \cdots, \alpha_j x_j \tag{3}$$

For count data,ε random error from the equation (1) often follows a Poisson distribution (Zar, 1999), and response variable y has nonnegative whole integers, also maximum likelihood techniques are applied to assess the best-fitted model. If the variance of observed y is greater than the expected value of$y$, over-dispersion occurs. According to Agresti (1996), if the over-dispersed parameter is calculated and multiplied with estimated standard errors, over-dispersion can be explained. In terms of count models, NB distribution has an extra parameter to account for over-dispersion (Crawley, 1997).

The number of involved nodes is considered to be a count outcome discrete variable, the Poisson regression model is the most common technique employed to model such count data. The probability mass function of the Poisson distribution is given as:.

$$P(y_j, \lambda) = \frac{e^{-\lambda}\lambda^{y_j}}{y_j!} y_j = 0, 1, 2, 3 \tag{4}$$

where random variable y_j, is the count response and parameter λ is the mean, and also variance. Poisson distribution has conditions of independent events and equality of mean and variance (Consul and Famoye, 1992). Due to the limitation of equal mean and variance, the Poisson distribution is not appropriate to fit the observed involved lymph nodes count response, since the variance of the counts, which is the number of involved lymph nodes, is much larger than their mean.

A solution to this overdispersion can be solved by applying a gamma-Poisson mixture distribution, which is known as the NB distribution. Its probability mass function is given as:.

$$P(y_j,\lambda,\tau)=\frac{\Gamma(\tau^{-1}+y_j)}{\Gamma(\tau^{-1})\Gamma(y_j+1)}\left(\frac{\tau^{-1}}{\tau^{-1}+\lambda}\right)^{\tau^{-1}}\left(\frac{\lambda}{\lambda+\tau^{-1}}\right)^{y_j} \tag{5}$$

The mean and variance of the negative binomial distribution are$E(y)=\lambda\tau$, and Var(y) $=\lambda\tau(1+\lambda)$. τ is the dispersion parameter, if$\tau=0$, negative binomial approaches to the Poisson model (Hilbe, 2001).

2.3. Zero-inflated models

For a better fit, an over-dispersed model that incorporates excess zeros is divided into two types, true and false zeros (Cheung, 2002), via zero-inflated models. True zeros are included in the study, which is part of the natural process, classified into structural and random zeros. False zeros are occurred due to observers' poor experience, caused due to sampling errors or errors in the experimental design (Tang et al., 2018). In breast cancer, patients' study excess zeros are assessed because of that group of patients who are "not-at-high risk" during the observation period, or who are "at-risk". For example, the number of lymph nodes involvement is an important factor in breast cancer prognostic and prevention research, but at a specific time some patients may not involve any lymph nodes, but later the chances of lymph nodes involvement may increase. It is also possible that there would be zero number of involved lymph nodes at a diagnostic stage, but still, that group may be at high risk. The negative nodal status is divided into two groups of patients, one with a very low-risk of involved nodes (structural zeros), and the other with a high-risk of involved nodes (random zeros). Zero-inflated models are used to account for overdispersion due to excess of zeros and unobserved heterogeneity among women diagnosed with breast cancer as a primary disease. Under zero-inflated modelling techniques, true zeros are described through logistic regression and false zeros via the zero-inflated part of the count model.

Zero-inflated models add additional probability mass to the outcome of excess zeros. It yields two states mixture distribution with PMF of ZIP model, which is given by:.

$$Pr(y_j,\lambda_j)=\begin{cases}\pi_j+(1-\pi_j)e^{-\lambda_j}, & y_j=0,0\le\pi\le1\\ \frac{(1-\pi)e^{-\lambda_j}(\lambda_j)^{y_j}}{y_{j!}}, & y_j\ge1\end{cases} \tag{6}$$

The ZINB distribution is used to account for both over-dispersion and excess of zeros. For extra zeros, it gives weight π, while $(1-\pi)$ weight is assigned to the negative binomial distribution, where a range of π is 0 to1. The mixture form of ZINB distribution can be written as:.

$$Pr(y_j,\lambda_j,\tau)=\begin{cases}\pi_j+(1-\pi_j)(1+\tau\lambda_j)^{-\tau^{-1}}, & 0<\pi<1\\ (1-\pi_j)\frac{\Gamma\left(y_j+\frac{1}{\tau}\right)(\tau\lambda_j)^{y_j}}{y_{j!}\left(\frac{1}{\tau}\right)1+\tau\lambda^{y_j+\frac{1}{\tau}}}, & y_j>\tau\end{cases} \tag{7}$$

where $\alpha\ge0$ is an over-dispersion parameter.

2.4. Hurdle models

The hurdle count models are two-part models. In the first part, zeros are modelled through logistic regression, and in the second part, the positive counts are explained through a zero-truncated Poisson or negative binomial distribution. HP model addresses, excess zeros in the first part, and truncated positive outcomes in the second part via zero-truncated Poisson distribution. The HP model can be written as:.

$$\begin{aligned}&\Pr(y_j=0)=1-\pi,\quad 0\le\pi\le1\\ &\Pr(Y=y_j)=\pi\frac{\exp(-\lambda_j)\lambda_j^{y_j}}{y_j!},\quad \lambda\ge0;y_j=1,2,3\end{aligned} \tag{8}$$

here, λ is the mean of the Poisson model.

The HNB is appropriate to model the data which exhibit over-dispersion due to only excess zeros (Cameron and Trivedi, 1999; Chipeta et al., 2014), so it does not account for unobserved heterogeneity which exists in our breast cancer data. The HNB model is given as:.

$$\Pr(y=0)=1-\pi,0\le\pi\le1$$

$$\Pr(Y=y)=\frac{\pi}{1-\left(\frac{\tau}{\lambda+\tau}\right)^{\tau}}\frac{\Gamma(y+\tau)}{\Gamma(\tau)y!}\left[\frac{\lambda}{\lambda+\tau}\right]^{y}\left[\frac{\tau}{\lambda+\tau}\right]^{\tau},\quad \tau,\lambda>0;y=1,2,3,\cdots \tag{9}$$

here, mean is λ and variance is $\lambda(1+\frac{\lambda}{\tau})$.

2.5. Model assessment and evaluation

Comparison of fitted models is done via measures of fit, which describe the performances of fitted models for a given data set, the good model is selected, based on log-likelihood, the Akaike information criteria (AIC), and the Bayesian information criteria (BIC).

$$AIC=-2(log-likelihood)+2(df) \tag{10}$$

$$BIC=-2(log-likelihood)+n(df) \tag{11}$$

where, df represents degrees of freedom of the fit, and n is the total number of observations in the data. The AIC criteria use penalize function as if add a variable, sampling variability also increases, and the BIC criteria impose a stronger penalty in the inclusion of additional variables to the model, a lower AIC and BIC values indicate that the model is to better fit for the understudy data (Vrieze, 2012; Pan, 2001). According to Vrieze (2012), model selection criteria between AIC and BIC depends upon the complexity of the true model. An information criteria difference, which is less than 4 shows indifference between two models, between 4 and 10 differences indicate that one model is moderately superior to the other, and a difference greater than 10 suggests that, one model is, in reality better than the other (Chipeta et al., 2014). For statistical analyses, the level of significance was chosen at 5%, and all modelling frameworks and analyses were carried out using Political Science Computational Laboratory (PSCL) package (Jackman, 2008) in R statistical software (The R Foundation for Statistical Computing, Version 3.6.2.).

2.6. Ethical approval

The study was approved by the Advanced Studies and Review Board, University of the Punjab, Lahore (Pakistan). After that, the letter of support written by the departmental head was submitted to the selected hospital. Prior to data collection, written consent was obtained from the head of the oncology department and con-

fidentiality was maintained by coding, from data collection to analysis. No written consent was needed because the analysis is based on routinely collected data, for which the hospital has already informed patients.

3. Results

3.1. Sample characteristics and lymph node involvement

The analysis is based on 5196 female patients with breast cancer, of which more than half (54.5%) were invasive ductal carcinoma. The median age at diagnosis was 48 years. The patients were almost equally distributed among the histological grades (I, II, III). About half of the patients were ER-positive (51.2%), PR-positive (51.0%), and Her2 positive (53.0%). The majority of the patients (70.2%) had a tumor size between 2 and 4.9 cm (Table 1).

Fig. 1 shows that a large proportion of individuals, i.e. overall, 2406 breast cancer patients (46,3%) had no lymph node involved at the diagnostic stage. There was overwhelming evidence of over-dispersion, which was confirmed by the presence of excess zeros (Fig. 1).

3.2. Model comparison

The comparison of models is presented in Table 2, using the values from the AIC and BIC for assessment basis. Although the zero-inflated negative binomial (ZINB) has a superiority over the hurdle negative binomial in terms of small AIC and BIC (AIC = 16,559, BIC = 16,710), in terms of zero-capturing (2,406) the hurdle negative binomial (HNB) model showed good performance (AIC = 16,587, BIC = 16,737). The difference between results of (AIC and BIC values) ZINB and HNB models is greater than 10, so the best model to fit the understudy data is ZINB, as it has the lowest AIC and BIC values.

The Poisson count model was not appropriate for this data set, because it only captured 1,262 numbers of zeros, same is with the NB model which captured 1,335 zeros out of a total 2,406. ZIP and ZINB were much better in capturing the zero counts 2,395 and 2,393 respectively. The best models to capture zeros were HP and HNB, both captured 2,406 zeros which were equal to the observed number of zeros (Table 3).

Also, it is important to consider that all patients were at risk of nodes involvement, so due to sampling zeros, inflated models are technically suitable to predict nodes involvement frequency among women diagnosed with primary breast cancer. It is important to be noted that the ZIP and ZHP models account for overdispersion due to excess zeros, but if overdispersion exists due to unobserved heterogeneity or progressive dependency in nodal involvement data, ZINB and HNB models give a better fit.

After a comparison of models, the ZINB model was used as the best-fitted model to count lymph nodes data in primary breast cancer patients and determination of factors that contributed to involved lymph node status.

Table 1
Descriptive statistics of 5,196 patients with breast cancer.

	n (%)
Tumor type	
IDC	2,832 (54.5%)
Other	2,364 (45.5%)
Baseline age (inyears)	
≤35	736 (14.2%)
36–45	1,092 (21.0%)
≥46	3,368 (64.8%)
Tumor grade	
I	2,021 (38.9%)
II	1,618 (31.1%)
III	1,557 (30.0%)
Estrogen receptor (ER)	
Positive	2,662 (51.2%)
Negative	2,534 (48.8%)
Progesterone receptor (PR)	
Positive	2,652 (51.0%)
Negative	2,544 (49.0%)
Human epidermal growth factor receptor 2 (Her2)	
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%)

3.3. Modelling and interpreting main effects

The final model ZINB, accounts for excess zeros count response data, having a mean number of involved nodes $((1-\pi)\lambda)$, and variance$(1-\pi)\lambda(1+\pi\lambda+\frac{\lambda}{\tau})$.

Table 4 provides the estimates of regression coefficients corresponding to various factors for the ZINB model with $a5\%$ level of significance. The NB (count) part of the ZINB model exhibits the risk of a greater number of lymph nodes, given that women are in a high-risk group. It is noted that patients of tumor grade II $(OR = 1.002,\ 95\%CI:\ 0.944 - 1.064)$ and III $(OR = 1.323, 95\%CI: 1.248 - 1.402)$ had a higher risk of having more involved lymph nodes as compared to grade I patients. ER-negative $(OR = 0.951,\ 95\%CI:\ 0.913 - 0.993)$, and PR-negative $(OR = 0.897,\ 95\%CI:\ 0.856 - 0.939)$ patients had a lower risk of having a greater number of nodes than ER and PR-positive patients. Results show that greater tumor size $2 - 4.9\ cm$ $(OR = 2.068,\ 95\%CI:\ 1.836 - 2.329)$ and $\geq 5cm$ $(OR = 5.230,\ 95\%CI:\ 4.625 - 5.913)$ have a higher likelihood of having a larger number of involved axillary nodes than $\leq 1.9cm$. Baseline age and Her2 status have not been significantly associated with nodal status.

Table 4 also contains ORs from the logistic part of the ZINB model, this part shows the probability of negative nodal status, given that breast cancer patients are in a low-risk group. Women of tumor type other than ductal carcinoma $(OR = 6.868,\ 95\%CI:\ 5.632 - 8.376)$ had a greater chance to exist in the negative nodal status group. Patients of tumor grade I had more chances of having no lymph node involvement. ER, PR and Her2-positive women significantly increase the likelihood of not having any number of axillary lymph nodes involvement at the initial stage of breast cancer. Women who had tumor size $\leq 1.9cm$ were more likely to not have any number of positive lymph node than those who had higher tumor size $2 - 4.9cm$ $(OR = 0.496, 95\%CI: 0.394 - 0.626)$ and $\geq 5cm(OR = 0.036,\ 95\%CI: 0.022 - 0.059)$. Baseline age has no significant impact on the negative nodal status.

4. Discussion

Breast cancer, a commonly diagnosed malignancy in females, represents a major public health issue worldwide (Barnard et al., 2015). Previous studies have shown a large absolute number 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 (Barnard et al., 2015). 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

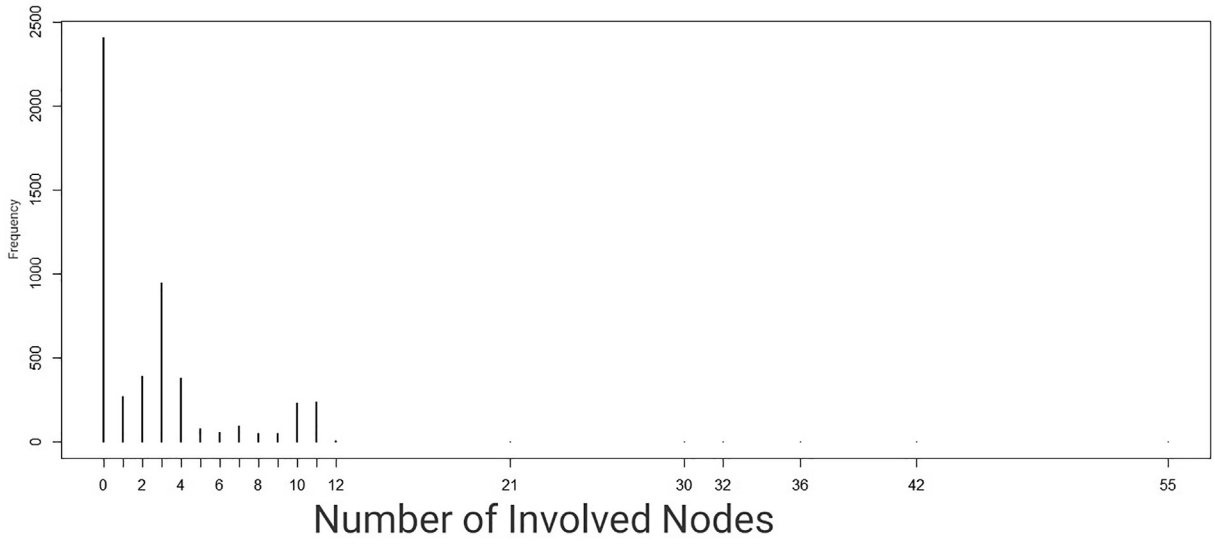

Fig. 1. Frequency of number of nodes.

Table 2
Model assessment for all models.

Model selection criterion	Poisson	NB	ZIP	ZINB	Hurdle P	Hurdle NB
df (degree of freedom)	11	12	22	23	22	23
Log-likelihood	−10378	−9375	−8307	−8257	−8318	−8271
AIC	20,778	18,774	16,658	16,559	16,680	16,587
BIC	20,849	18,851	16,803	16,710	16,824	16,737

Table 3
Zero count capturing in the understudy models.

Observed	Poisson	NB	ZIP	ZINB	Hurdle P	Hurdle NB
2406	1262	1607	2395	2393	2406	2406

clinical stage, greater tumor size, and metastasis (Barnard et al., 2015; Gann et al., 1999; Olivotto et al., 1998; Ravdin et al., 1994).

The presence or absence of auxiliary lymph nodes has been recognized as an important predictor of breast cancer risk. Studies have shown node-positive patients had lower survival rates than node-negative ones (Chua et al., 2001; Fisher et al., 1983). Furthermore, it is studied that a higher number of positive lymph node involvements contributes to an increased risk of complications (Chua et al., 2001; Fisher et al., 1983). 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 (Harden et al., 2001; Sakorafas et al., 2000). The research applied statistical distributions for involved lymph nodes in breast cancer, this study highlighted the problem of overdispersion due to temporal dependency in axillary involved nodes (Guern and Vinh-Hung, 2008).

Data involving the number of lymph nodes often contains surplus zeros, which indicates overdispersion in the data, therefore such data must be fitted by zero-inflated and zero-hurdle models. To estimate false zeros in the axillary involved lymph node data, it is to be noted that some negative nodal status might be observed among women who were at a "low risk" group of breast cancer and some among women who were at a "high risk" group of breast cancer. It is because not all women possess an equal intensity of breast cancer while having different tumor types, tumor grade, the status of ER, PR, Her2, and tumor size. With this logical consideration, the ZINB model is employed as a final model. The better fit of the ZINB model over the HNB model suggests that overdispersion is due to unobserved heterogeneity among women regarding the intensity of breast cancer and a larger number of negative nodal status as well.

In this study, not only fitted and compared several count models to investigate the number of involved nodes in primary breast cancer patients, but we also explained the significance of applying zero-inflated models in case when there exist both true and false zeros. Results of data analysis recommended the effectiveness of zero-augmented (zero-inflated and hurdle) models as compared to generalized linear (Poisson and negative binomial) models. Due to the excess of zeros and over-dispersion examined in this study, zero-augmented negative binomial models (ZINB and HNB) performed better than zero-augmented Poisson models (ZIP and HP). The ZINB and HNB models are similar in identifying factors associated with the number of involved lymph nodes. The ZINB model has been found to provide the best fit for modelling the involved lymph nodes data as a response variable and patients' age, tumor type, tumor grade, molecular subtypes, and tumor size as the explanatory variables in primary breast cancer patients. Our model selecting logic is the same as the results presented in the articles (Rose et al., 2006; Baughman, 2007), it is also suggested that model selection should be based on study objectives.

The best model ZINB was used to determine significant factors, which influence the number of involved lymph nodes in breast cancer patients. Women with higher tumor grades (II and III), estrogen and progesterone receptors positive, and higher tumor

Table 4
Estimates for ZINB model for number of lymph nodes in breast cancer study.

Parameter	Lymph nodesOR (95% CI)	Zero-inflation portion Lymp nodes OR (95% CI)
Intercept	1.626 (1.416–1.867)	3.538 (2.527–4.952)
Tumor type		
IDC	1	1
Other	1.010 (0.955–1.068)	6.868 (5.632–8.376)
Age (in years)		
≤35	1	1
36–45	0.995 (0.926–1.070)	1.089 (0.838–1.417)
≥46	1.020 (0.958–1.085)	1.040 (0.831–1.302)
Tumor grade		
1	1	1
11	1.002 (0.994–1.064)	0.374 (0.309–0.454)
111	1.323 (1.248–1.402)	0.453 (0.372–0.550)
Estrogen receptor		
Positive	1	1
Negative	0.951 (0.913–0.993)	0.544 (0.460–0.642)
Progesterone receptor		
Positive	1	1
Negative	0.897 (0.856–0.939)	0.276 (0.229–0.334)
Her2.neu receptor		
Positive	1	1
Negative	0.969 (0.928–1.014)	0.429 (0.363–0.508)
Tumor size (in cm)		
≤1.9	1	1
2–4.9	2.068 (1.836–2.329)	0.496 (0.394–0.626)
≥5	5.230 (4.625–5.913)	0.036 (0.022–0.059)

size are factors contributing to a greater number of positive involved lymph nodes. Age at diagnosis does not have any significant impact on nodal involvement in primary breast cancer.

4.1. Limitations

Some limitations may be noted. First, the use of a single case study may be viewed as a limitation; a simulation study can be conducted to strengthen our conclusions. Second, we were not able to account a longitudinal assessment that may reveal other aspects related to ‘‘high risk” and ‘‘low risk” groups of understudy data. Apart from these future tasks, this study is trying to fill the statistical modelling 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 hospitals where patients come from all over Pakistan.

4.2. Conclusions

Zero-inflated models assume zeros can be both true or false zeros, such zeros are estimated by binary and count components, while hurdle models (HP and HNB) assume that all zeros are true zeros and all patients belong to the same high-risk group. We believe that our study successfully quantified the ‘‘high and low risk” breast cancer patients by incorporating time-independent covariates which are associated with the presence of involved lymph nodes, so the zero-inflated negative binomial model was the best choice. Also, we applied hurdle models, which have successfully demonstrated the advantage of fitting count nodal data, it has two components; a binary logit model for positive counts, and a negative binomial model for truncated below at 1.

In short, the conclusion is, this paper provides the evidence to support that involved node count data at primary breast cancer are rightly skewed with excess zeros, so should be modeled by the zero-augmented negative binomial models. Between ZINB and HB models, the ZINB model is considered to be the best model for describing the number of involved nodes in primary breast cancer patients.

5. Ethical approval and consent to participate

According to the Ethical Guidelines for Epidemiologists (IEFEGE) and the regulations of the ethics committee located at the Advanced Studies and Review Board, University of the Punjab Lahore (Pakistan), no ethics approval is needed, because the analysis is based on routine data. At data collection, all patients provided written informed consent.

6. Consent for publication

Not applicable.

7. Availability of data and materials

Data is available from corresponding author upon reasonable request.

8. Authors’ contributions

ML conceived the original idea of the study, designed the study, analyzed the data and drafted the manuscript. SK supervised the whole study design. NZ has been responsible for data acquisition. SK and FF revised it critically for important intellectual content. All authors approved the final version of the manuscript.

Funding

The work was supported by the Higher Education Commission Pakistan under grant No. 46-2SS2- 123 awarded to the first author. The funder had no role in study design, in the collection, analysis and interpretation of data, in the writing of the report, and in the decision to submit the article for publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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 he put to organize patients' records so systematically.

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

References

Afifi, A.A., Kotlerman, J.B., Ettner, S.L., Cowan, M., 2007. Methods for improving regression analysis for skewed continuous or counted responses. Ann. Rev. Public Health. 28, 95–111. https://doi.org/10.1146/annurev.publhealth.28.082206.094100.

Agresti, A., 1996. An Introduction to Categorical Data Analysis. John Wiley and Sons, New York.

Barnard, M.E., Boeke, C.E., Tamimi, R.M., 2015. Established breast cancer risk factors and risk of intrinsic tumor subtypes. Biochim. Biophys. Acta. 1856 (1), 73–85. https://doi.org/10.1016/j.bbcan.2015.06.002.

Baughman, L.A., 2007. Mixture model framework facilitates understanding of zero-inflated and hurdle models for count data. J. Biopharm. Stat. 17 (5), 943–946. https://doi.org/10.1080/10543400701514098.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., Kirchner, U., 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. J. R. Statis. Soc. A. 162 (2), 195–209.

Cameron, A.C., Trivedi, P.K., 1999. Essentials of Count Data Regression (Chapter 15). A Companion to Theoretical Econometrics. Blackwell Publishing Ltd., Malden.

Cheung, Y.B., 2002. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat. Med. 21, 1461–1469. https://doi.org/10.1002/sim.1088.

Chipeta, M.G., Ngwira, B.M., Simoonga, C., Kazembe, L.N., 2014. Zero adjusted models with applications to analysing helminths count data. BMC Res. Notes. 7, 856. https://doi.org/10.1186/1756-0500-7-856.

Chollet-Hinton, L., Anders, C.K., Tse, C.-K., Bell, M.B., Yang, Y.C., Carey, L.A., Olshan, A. F., Troester, M.A., 2016. Breast cancer biologic and etiologic heterogeneity by young age and menopausal status in the Carolina Breast Cancer Study: a case-control study. Breast Cancer Res. 18 (1). https://doi.org/10.1186/s13058-016-0736-y.

Chua, B., Ung, O., Taylor, R., Boyages, J., 2001. Frequency and predictors of axillary lymph node metastases in invasive breast cancer. A. N. Z. J. Surg. 71 (12), 723–728. https://doi.org/10.1046/j.1445-1433.2001.02266.x.

Consul, P.C., Famoye, F., 1992. Generalized Poisson regression model. Commun. Statist. – Theory Methods. 2 (1), 89–109. https://doi.org/10.1080/03610929208830766.

Crawley, M.J., 1997. GLIM for Ecologists. Blackwell Science, Oxford.

Fisher, B., Bauer, M., Wickerham, D.L., Redmond, C.L.K., Fisher, E.R., 1983. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. Cancer. 52 (9), 1551–1557. https://doi.org/10.1002/1097-0142(19831101)52:9<1551::aid-cncr2820520902>3.0.co;2-3.

Gann, P.H., Colilla, S.A., Gapstur, S.M., Winchester, D.J., Winchester, D.P., 1999. Factors associated with axillary lymph node metastasis from breast carcinoma descriptive and predictive analyses. Cancer. 86 (8), 1511–1518. https://doi.org/10.1002/(sici)1097-0142(19991015)86:8<1511::aid-cncr18>3.0.co;2-d.

Gilthorpe, M.S., Frydenberg, M., Cheng, Y., Baelum, V., 2009. 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. 28 (28), 3539–3553. https://doi.org/10.1002/sim.3699.

Guern, A.S., Vinh-Hung, V., 2008. Distribution statistique des ganglions lymphatiques axillaires envahis lors du cancer du sein [Statistical distribution of involved axillary lymph nodes in breast cancer]. Bull. Cancer. 95 (4), 449–455. https://doi.org/10.1684/bdc.2008.0620.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 56 (4), 1030–1039. https://doi.org/10.1111/j.0006-341x.2000.01030.x.

Harden, S.P., Neal, A.J., Al-Nasiri, N., Ashley, S., Quercidella, R.G., 2001. Predicting axillary lymph node metastases in patients with T1 infiltrating ductal carcinoma of the breast. Breast. 10 (2), 155–159. https://doi.org/10.1054/brst.2000.0220.

Hilbe, J.M., 2001. Negative Binomial Regression. Cambridge University Press, Cambridge.

Hur, K., Hedeker, D., Henderson, W., Khuri, S., Daley, J., 2002. Modeling clustered count data with excess zeros in health care outcomes research. Health Serv. Outcomes Res. Methodol. 3, 5–20. https://doi.org/10.1023/A:1021594923546.

Jackman, S., 2008. Classes and methods for R developed in the political science computational laboratory, Stanford University. Stanford, California: Department of Political Science, Stanford University. R package version 0.95; http://CRAN.R-project.org/package=pscl.

Lambert, D., 1992. Zero-inflated poisson regression with an application to defects in manufacturing. Technometrics 34 (1), 1–14. https://doi.org/10.2307/1269547.

Lee, A.H., Wang, K., Yau, K.K.W., 2001. Analysis of zero-inflated Poisson data incorporating extent of exposure. Biometrical J. 43 (8), 963–975. https://doi.org/10.1002/1521-4036(200112)43:8<963::AID-BIMJ963>3.0.CO;2-K.

Lee, A.H., Wang, K., Scott, J.A., Yau, K.K., McLachlan, G.J., 2006. Multi-level zero-inflated Poisson regression modeling of correlated count data with excess zeros. Stat. Methods Med. Res. 15 (1), 47–61. https://doi.org/10.1191/0962280206sm429oa.

Moloas, M., Lesaffre, E., 2010. Hurdle models for multilevel zero-inflated data via h-likelihood. Stat Med. 29 (30), 3294–3310. https://doi.org/10.1002/sim.3852.

Mullahy, J., 1986. Specification and testing of somemodified count data models. J Econometrics. 33 (3), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3.

Olivotto, I.A., Jackson, J.S.H., Mates, D., Andersen, S., Davidson, W., Bryce, C.J., Ragaz, J., 1998. Prediction of axillary lymph node involvement of women with invasive breast carcinoma a multivariate analysis. Cancer. 83 (5), 948–955. https://doi.org/10.1002/(SICI).

Pan, W., 2001. Akaike's information criterion in generalized estimating equations. Biometrics. 57 (1), 120–125.

Ravdin, P.M., De Laurentiis, M., Vendely, T., Clark, G.M., 1994. Prediction of axillary lymph node status in breast cancer patients by use of prognostic indicators. J. Natl. Cancer Inst. 86 (23), 1771–1775. https://doi.org/10.1093/jnci/86.23.1771.

Rose, C.E., Martin, S.W., Wannemuehler, K.A., Plikaytis, B.D., 2006. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J. Biopharm. Stat. 16 (4), 463–481. https://doi.org/10.1080/10543400600719384.

Sakorafas, G.H., Tsiotou, A.G., Balsifer, B.M., 2000. Axillary lymph node dissection in breast cancer: current status and controversies, alternative strategies and future perspectives. Acta Oncol. 39 (4), 455–466. https://doi.org/10.1080/028418600750013366.

Suissa, S., Blais, L., 1995. Binary regression with continuous outcomes. Stat. Med. 14 (3), 247–255. https://doi.org/10.1002/sim.4780140303.

Tang, W., He, H., Wang, W.J., Chen, D.G., 2018. Untangle the structural and random zeros in statistical modelings. J. Appl. Stat. 45 (9), 1714–1733. https://doi.org/10.1080/02664763.2017.1391180.

Vrieze, S.I., 2012. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol. Methods. 17 (2), 228–243. https://doi.org/10.1037/a0027127.

Winkelmann, R., Zimmermann, K.F., 1995. Recent developments in count data modelling: theory and application. J. Econ. Surv. 9 (1), 1–24. https://doi.org/10.1111/j.1467-6419.1995.tb00108.x.

Yau, K., Wang, K., Lee, A., 2003. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical J. 45 (4), 437–452. https://doi.org/10.1002/bimj.200390024.

Yesilova, A., Kaydan, M.B., Kaya, Y., 2010. Modeling insect-egg data with excess zeros using zero-inflated regression models. Hacettepe J. Math. Stat. 39 (2), 273–282.

Zar, J.H., 1999. Biostatistical Analysis. Prentice-Hall, Upper Saddle River.

Zorn, C.J.W., 1996. Evaluating zero-inflated and hurdle Poisson specifications. Midw. Polit. Sci. Assoc., 1–16