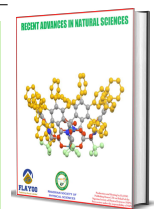
Published by Nigerian Society of Physical Sciences. Hosted by FLAYOO Publishing House LTD

Recent Advances in Natural Sciences

Journal Homepage: https://flayoophl.com/journals/index.php/rans



Zero-inflated and hurdle models with an application to the number of involved axillary lymph nodes in patients with breast cancer in Zimbabwe: A bootstrap resampling validation approach

Bester Saruchera[a,*], Oliver Bodhlyera[a], Henry Mwambi[a], Ntokozo Ndlovu[b]

[a]*School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg,3209, South Africa*
[b]*Department of Oncology, Faculty of Medicine and Health Sciences, University of Zimbabwe, Harare, Zimbabwe*

ARTICLE INFO

Article history:
Received: 02 December 2024
Received in revised form: 27 January 2025
Accepted: 31 January 2025
Available online: 27 February 2025

Keywords: Breast cancer, Axillary lymph nodes, Count regression models, Bootstrapping resampling

DOI:10.61298/rans.2025.3.1.137

ABSTRACT

Breast cancer, increasingly prevalent in Zimbabwe, underscores the need to understand the involved axillary nodal status in diagnosed patients for assessment of disease severity and its potential progression. This study was undertaken to investigate factors influencing the number of axillary lymph nodes in breast cancer patients by identifying the best fitting count regression model, validated through bootstrap resampling. A retrospective analysis using hospital-based data for patients diagnosed with breast cancer at one of the two major referral hospitals in Zimbabwe was applied. We evaluated and compared count regression models – Poisson with Negative Binomial (NB), Zero-Inflated Negative Binomial (ZINB), Zero-Inflated Poisson (ZIP), Hurdle Negative Binomial (HNB) and Hurdle Poisson (HP) which are efficient in handling over-dispersed count data to investigate the various risk factors associated with involved axillary lymph nodes. Covariates included age, tumor size, tumor grade, estrogen receptor status, progesterone receptor status and HER2 status. Model diagnostics were assessed using Aikake Information Criterion and Bayesian Information Criterion. The ZINB and HNB models outperformed other models, with the HNB model demonstrating consistency across bootstrap-resampled datasets. Bootstrap resampling validated the reliability of model estimates, addressing biases caused by small sample sizes. Age was significantly associated with the zero-inflation component of the HNB model. This study highlights the importance of selecting appropriate count regression models for analyzing medical data and demonstrates the utility of integrating bootstrap resampling to ensure robust statistical inference. The findings provide actionable insights for therapy planning and resource allocation.

1. INTRODUCTION

Breast cancer is the second most common cancer in Zimbabwean women after cervical cancer [1]. According to International Agency for Research on Cancer (IARC) [2], breast cancer

*Corresponding author
e-mail: bester524@gmail.com (Bester Saruchera)

accounted for 17.1% of cancers in 2018 in Zimbabwe, and globally in 2020 breast cancer had the highest incidence at 11.7% compared to other cancers. With a staggering projection of an increase of breast cancer incidence in Zimbabwe from 1,886 in 2018 to 4,185 in 2040 by the World Health Organisation, more advances need to be made in diagnostic and treatment strategies.

Axillary lymph node involvement is an essential component in the staging of breast cancer. In the TNM staging system, T represents the size and extent of the tumor which can be assessed clinically or through imaging techniques whilst M signifies the presence or absence of distant metastases [3]. Our study, however, focuses on the N component, representing lymph nodes involvements which can be determined clinically by examination and pathologically through processes that include needle biopsy axillary surgery. Axillary nodal status is one of the most important prognostic factors in breast cancer which has also become an important predictive tool to adapting systemic therapies [4]. The lymph node status gives guidance to therapeutic strategies and is a measure of prognosis in patients especially those with distant metastases [5]. If a tumor has not spread to the lymph nodes, chances of the cancer being treatable are high.

With the number of involved axillary lymph nodes, which is a count variable not normally distributed within the population of individuals diagnosed with breast cancer, the Poisson Regression model (PRM) would provide an appropriate analysis where the dependent count variable may be explained by the covariates. The PRM however, has a restrictive assumption of equality of the mean and variance of the outcome variable, conditional on the covariates, whilst the Negative Binomial (NB) assumes an over-dispersion of the outcome variable. The NB model, also known as the Poisson-gamma, is a generalization of the Poisson model with a two parameter distribution that relaxes the restrictive assumption made in Poisson regression [6].

Zero count data and over-dispersion are often ubiquitous in medical health investigations. Lambert [7], proposed the Zero Inflated Poisson (ZIP) as a model for count data with extra zeros, as a two part model assuming that with probability p , 0 is the only possible observation and with probability 1-p, any positive integer is observed. Greene [8], introduced the Zero-Inflated Negative Binomial (ZINB) with capacity to efficiently model count outcomes with over-dispersion and excess zeros. Zero inflated count models assume that the data is a mixture of two separate data generation processes, the first one generating zeros only and the second one being either a Poisson or a NB data generating process [9]. The zero observations have two different origins: "structural" and "sampling". Hurdle Poisson (HP) and Hurdle Negative Binomial (HNB) models have also been developed to model zero-inflation where the Poisson or NB models are unrealistic [10, 11]. These models which assume that zero counts in the outcome are generated from a different process than the positive counts can be an alternative if the positive counts have extra overdispersion [6].

Recent developments have been proposed to address traditional Poisson regression models limitations. For instance, Adesina [12] employed a Bayesian multi-level framework using the No-U-Turn Sampler to sample from posterior distributions and demonstrated its effectiveness on both over- and under-dispersed data simulated from a discrete Weibull distribution. Their findings underscored the superiority of the geometric model over other alternatives for handling over-dispersed data. Maxwell *et al.* [13] compared generalized poisson regression to traditional and advanced count models (Poisson, NB, and Conway-Maxwell-Poisson) in the context of road traffic crash data and found that Generalized Poisson had the smallest AIC and BIC values, indicating superior model fit. This reinforces the importance of using tailored models for count data to achieve robust and reliable results.

Count models have also been applied to medical outcomes, including in counts of involved lymph nodes in breast cancer. Swain *et al.* [14], Feng *et al.* [15], and Liaqat *et al.* [16] utilized such models to analyze lymph node involvement in breast cancer, while Abu *et al.* [17] applied them to health services utilization. Nketia *et al.* [18] demonstrated their utility in studying schistosomiasis, and Yirga *et al.* [19] applied similar approaches in HIV-related studies, specifically for CD4 count data. Pavlicova *et al.* [20] applied the models to assess the effect of an HIV-risk reduction intervention on unprotected sexual occasions in a clinical trial, ultimately identifying the ZINB model as providing the best fit. These applications highlight the versatility of zero-inflated and hurdle models in addressing over-dispersion and excess zeros in medical data, aligning closely with the objectives of this study. Despite their utility, inconsistencies often arise in findings due to variations in model specifications and data characteristics. For example, some studies reported that ZINB models outperform other count models, while others found ZINB and hurdle models indistinguishable in terms of goodness-of-fit measures [15, 16, 18, 20–24]. These variations emphasize the need for further research to clarify the relative performance of these models under different data conditions, a focus that this study aims to address.

Moreover, while zero-inflated and hurdle models are well-documented approaches for analyzing over-dispersed count data, limited research has focused on assessing the stability of model selection procedures. Integrating bootstrap resampling for validating model coefficients offers a promising yet underexplored alternative for improving model reliability. Although count models provide valuable insights into associations with covariates, their coefficient estimates are often sensitive to sample variability, a challenge that becomes especially pronounced in studies with limited observations. This underscores the need for robust methodological approaches that enhance the stability and reliability of statistical inference, particularly in resource-constrained settings. Larger sample sizes generally provide more statistical power, allowing for more precise and reliable estimates of model parameters [25]. However, in resource-constrained contexts, obtaining large sample sizes can be challenging due to limited resources, access to data, and other logistical constraints. This limitation often results in highly variable parameter estimates, potentially undermining the reliability of conclusions. To address this, our study introduces a novel approach by integrating bootstrap resampling, a technique introduced by Efron [26], with count models to enhance the robustness and reliability. The resampling process generates multiple bootstrap samples from original observations with replacement, allowing for the estimation of variability and stability in model coefficients. This approach yields more reliable confidence intervals and mit-

igates biases caused by small sample sizes, which are rarely addressed in existing studies. Although similar resampling techniques have been applied in other fields, such as those demonstrated by Troung *et al.* [27] and Sillabutra *et al.* [28], their application to count regression models in medical research remains limited. This methodological advancement provides a valuable framework for robust analysis in resource-constrained settings, ensuring reproducibility and applicability across similar contexts.

Our overall objective was to identify the best-fitting model to analyze the risk factors associated with axillary nodal involvement, utilizing real life, distinct hospital based data collected at one of the two public referral oncology centres in Zimbabwe. By applying the innovative integration of bootstrap resampling in the framework of count models to this unique setting, we provide reliable insights into the factors influencing axillary lymph node involvement in breast cancer patients. This dataset enables us to address specific challenges faced in a Sub-Saharan country, such as limited healthcare resources and access to timely diagnosis and treatment, contributing to a deeper understanding of breast cancer prognosis in similar settings. By improving the precision of model estimates, our approach has significant implications for clinical decision-making and policy formulation. Specifically, the findings aim to inform resource allocation strategies and facilitate the identification of high-risk patients for targeted interventions. Ultimately, this research contributes to improved patient outcomes and offers a robust framework for advancing breast cancer care in resource-constrained environments.

2. MATERIALS AND METHODS

This section describes the data used in the study and the model building process. The mathematical formulation of the count regression models used is also discussed together with the model performance measures.

2.1. STUDY DESIGN

This retrospective study makes use of hospital data compiled from the records of 379 patients diagnosed with breast cancer between January 2015 and December 2019 at Parirenyatwa Radiotherapy Centre, which is one of the two referral oncology centres in Zimbabwe. Only 136 patients had their nodal count recorded. Covariates considered in this study were guided by the literature which included demographic and clinical variables: age, tumor size (*cm*), tumor grade (one, two and three), estrogen receptor (ER) status (positive / negative), progesterone receptor (PR) status (positive / negative) and Human Epidermal Growth factor Receptor 2 (HER2) status (positive /negative) [14, 16, 24]. The outcome count variable was the number of involved axillary lymph nodes. The study was approved by the Medical Research Council of Zimbabwe (MRCZ) and Joint Research and Ethics Committee (JREC) of the Parirenyatwa Group of Hospitals and University of Zimbabwe.

Figure 1 is the flowchart of the proposed steps in selecting the best count regression model. We visually examined the number of involved lymph nodes using a bar plot in Figure 2. Data preprocessing was conducted before statistical model building. The variables ER, PR and HER2 were label encoded and tumor grade was treated as an ordinal variable. Data normalization is an important step in data preprocessing as it improves model performance and training stability [29]. We applied the min-max normalization technique where the numeric values of age and tumor size were reduced to a scale between 0 and 1 using:

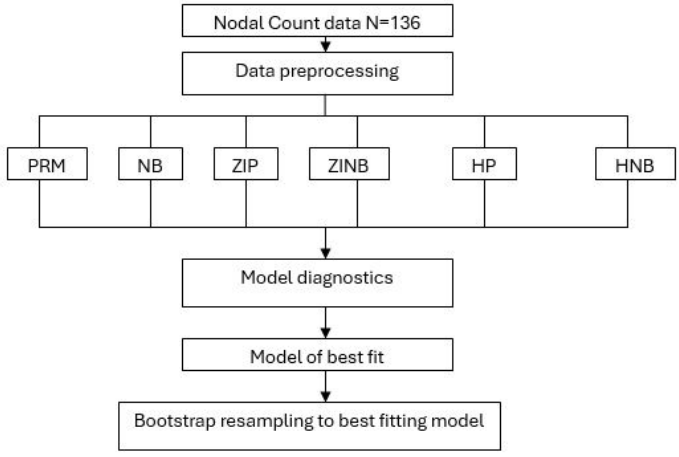


Figure 1. Flowchart for involved lymph nodal count model selection procedure. This illustrates the step-by-step process for selecting the appropriate model for analyzing involved lymph nodal counts. Each step is detailed to ensure clarity in the decision-making process, highlighting key criteria and decision points.

$$z_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)}, \tag{1}$$

where min and max are the minimum and maximum values of the mentioned numerical variables over the given data range.

2.2. MODELING FRAMEWORK

Count models allow regression analyses when the outcome variable of interest is a numerical count. The PRM was explored and checked for overdispersion. The NB, an alternative to the PRM [11], especially useful for data with overdispersion was then employed. In total, six, models, which represented various combinations of three types of generalized count regression models (one part, zero-inflated and hurdle) with one of the two distributions (PRM or NB) were fitted. The model parameters, derived by a log-likelihood estimator give the estimated effect of each covariate on the involved axillary nodal count data. The Pearson dispersion statistic which was used to identify over dispersion was computed and interpreted in all models. A dispersion parameter greater than 1 implied that the involved axillary nodal count data was more variable than would be expected under standard Poisson or negative binomial distribution, with a value less than 1 indicating that the nodal count data is less variable than expected. We interpreted the ZIP, ZINB and hurdle models based on the zero-inflation component and NB or Poisson component. Over and above the interpretation of the coefficients and dispersion parameters of the models, we evaluated the goodness of fit measures using the Akaike Information Criterion (AIC), Bayesian Information criterion (BIC) and AICc to determine the best fitting model. A $p < 0.05$ was considered as the statistical level of significance for the predictors used in all the models. Statistical analysis was done in R software.

2.2.1. Statistical models formulation

Count regression models employed in the study are described in detail below.

Poisson regression model (PRM):

The PRM is a Generalized Linear Model (GLM), the standard model for modelling count data with a non-negative integer valued dependent count variable [30]. In this study, Y_i denotes the count output variable which is number of axillary lymph nodes involved in the i^{th} women diagnosed with breast cancer: $y_i \in \{0, 1, 2, \ldots\}$. Given that Y_i is a counts variable, we assume that each observation y_i is from a Poisson distribution with mean μ_i (mean number of involved nodes): $Y_i \sim \text{Poisson}(\mu_i)$. The parameter μ_i controls the count rate in the i^{th} outcome. Thus, the PRM is derived from the assumption that the logarithm of the expected value of the outcome variable can be modeled as a linear combination of unknown parameters and is given by

$$\ln \mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq} = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij}, \tag{2}$$

or, equivalently

$$\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}} = e^{\beta_0 + \sum_{j=1}^{q} \beta_j x_{ij}}. \tag{3}$$

The probability mass function of the Poisson random variable with parameter μ_i is given by:

$$f(Y_i = y_i \mid x_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots \tag{4}$$

Assuming that the sample outcomes are independent, the joint likelihood function of sample outcomes is given by:

$$L(\mu_i; y_i, \ldots, y_n) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \tag{5}$$

and the log-likelihood function is given as:

$$\ln L(\mu_i; y_i, \ldots, y_n) = -\sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(y_i!) + \ln(\mu_i) \sum_{i=1}^{n} y_i. \tag{6}$$

Assuming a linear relationship between the log of the mean and the covariates, we have $E(y_i \mid x_i) = \mu_i = \exp(x_i'\beta)$ also expanded as $e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q}$. The intercept is β_0 and the regression coefficients for the covariates are $\beta_1, ..., \beta_q$, where q is the number of covariates in the model which represent the expected change in the log of the mean per unit change in the covariates. We thus can express the log-likelihood given as:

$$\ln L(\beta) = \sum_{i=1}^{n} \left(y_i x_i'\beta - \exp(x_i'\beta) - \ln(y_i!) \right). \tag{7}$$

We then optimized Eq. (7) iteratively up to convergence to a maximum value, that is until

$$\ln L(\beta) = \sum_{i=1}^{n} \left(y_i x_i'\beta - \exp(x_i'\beta) - \ln(y_i!) \right) = \max. \tag{8}$$

There are several optimization algorithms commonly employed for parameter estimation such as the Newton-Raphson (NR), Iterated Re-weighted Least Squares (IRWLS) also considered as Fisher Scoring and quasi-likelihood estimation [31, 32]. In this specific study, we utilized the IRWLS due to faster convergence and accuracy.

Overdispersion assessment using Pearson residuals The PRM has a restrictive property which requires the equality of the mean and variance of the outcome count variable. Over-dispersion arises when the variance of the response count variable is greater than the mean that is $Var[y_i|x_i] > E[y_i|x_i]$. Using the PRM in over-dispersed data leads to inefficient estimates and biased standard errors, most often, less than those inherent in the data [9]. However, real life medical data is often characterized with over-dispersion. Failure to correct for such bias in model standard errors, may lead to a higher risk of committing type I error more often than at the prescribed level of significance. We checked for over-dispersion using the ratio of the sum of the squared Pearson residuals to the residual degrees of freedom wherein a value greater than 1 indicates over-dispersion, and as such the NB becomes a better alternative to the PRM.

Negative binomial model (NB):

The NB distribution is a mixture of a family of Poisson distributions with gamma mixing weights [33]. The NB model can be considered as a generalization of the PRM that allows a variance higher than the corresponding mean. This model has commonly been used to analyse over-dispersed data in other studies [19, 24]. Structurally, the NB model adds a parameter which allows the conditional variance of the count variable to exceed the conditional mean. This parameter is among the parameters to be estimated from the likelihood of the data. We thus subsequently propose the negative binomial method where an extra (dispersion) parameter,φ, is introduced in the random structure $Y_i \sim \text{Poisson}(\mu_i)$ to control for overdispersion.

Model formulation:

In formulating the NB model, we start with the standard PRM for count data model generation which we adjust in the systematic structure resulting in the mean $\mu_i = \exp(x_i'\beta + \epsilon_i)$, where ϵ is the random error term assumed to be not correlated with the covariates $\mathbf{x_i}$. Defining the error term, ϵ_i to be $\log(\varphi_i)$, the adjusted random structure of the PRM becomes:

$$Y_i|\varphi_i \sim Poisson(\mu_i \varphi_i), \tag{9}$$

where φ_i is non negative and is assumed to follow a gamma distribution [6, 34, 35]: $\varphi_i \sim Gamma(a, b)$. In order for the mean of Y_i to remain unchanged as $E[Y_i] = \mu_i$, an assertion of the conditional expectation, $E[Y_i|\varphi_i] = \mu_i$ is obtained in the following steps:

Given:

$$E[Y_i] = E_{\varphi_i}[E[Y_i|\varphi_i]] = E_{\varphi_i}[\mu_i \varphi_i] = \mu_i E[\varphi_i] = \mu_i. \tag{10}$$

This implies that $E[\varphi_i] = 1$. With the assumption that φ_i follows a gamma distribution, it follows that $E[\varphi_i] = \frac{a}{b}$ and $Var(\varphi_i) = \frac{a}{b^2}$. Since we have $E[\varphi_i] = 1$, and by letting $\sigma^2 = Var(\varphi_i)$, then we have $a = b = \frac{1}{\sigma^2}$. This results in the 1-parameter Gamma-Poisson mixture model, that is, $\varphi_i \sim Gamma(\frac{1}{\sigma^2}, \frac{1}{\sigma^2})$. Having Y_i

taking discrete values with the conditional Poisson distribution

$$P(Y_i = y_i \mid \varphi_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \tag{11}$$

and φ_i following a gamma distribution:

$$f(\varphi_i, \frac{1}{\sigma^2}, \frac{1}{\sigma^2}) = \frac{(\varphi_i)^{\frac{1}{\sigma^2}-1} e^{-\frac{1}{\sigma^2}\varphi_i} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)}. \tag{12}$$

We evaluate the integral of the product of Poisson and the gamma distribution functions using:

$$P(Y_i = y_i) = \int_0^{+\infty} P(Y_i = y_i \mid \varphi_i) f(\varphi_i)\, d\varphi_i, \tag{13}$$

to obtain the following negative binomial (Poisson-Gamma) model:

$$f(Y_i \mid \varphi_i) = \frac{\Gamma\left(\frac{1}{\sigma^2} + y_i\right)}{y_i!\,\Gamma\left(\frac{1}{\sigma^2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}} \frac{\mu_i^{y_i}}{\left(\mu_i + \frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}+y_i}}. \tag{14}$$

Defining the dispersion parameter of φ_i as $\Phi = \frac{1}{\sigma^2}$ we have the probability function:

$$P(Y_i = y_i) = \frac{\Gamma(\Phi + y_i)}{y_i!\,\Gamma(\Phi)} (\Phi)^{\Phi} \frac{\mu_i^{y_i}}{(\mu_i + \Phi)^{\Phi+y_i}}. \tag{15}$$

Similarly, as in the PRM, from the systematic structure of the model,

$$\mu_i = \exp\left(x_i'\beta\right), \tag{16}$$

it follows that

$$\log(\mu_i) = x_i'\beta. \tag{17}$$

Lawless [36] applied one of the properties of the gamma distribution

$$\Gamma(\Phi + 1) = \Phi\Gamma(\Phi), \tag{18}$$

which results in the expression:

$$\frac{\log\Gamma(\Phi + y_i)}{\Gamma(\Phi)} = \sum_{j=1}^{y_i-1} \log(\Phi + j). \tag{19}$$

This leads to the negative binomial log-likelihood, parameterized in terms of β, which are the model coefficients expressed as:

$$\mathcal{L}(\Phi, \beta) = \sum_{i=1}^{n} \left[\left(\sum_{j=0}^{y_i-1} \log(\Phi + j)\right) - \log(y_i!) + \Phi\log(\Phi) + y_i(x_i'\beta) - (\Phi + y_i)\log\left(\Phi + \exp\left\{x_i'\beta\right\}\right)\right]. \tag{20}$$

The parameter Φ as introduced in Eq. (15) represents overdispersion. It is well noted that when $\Phi = 0$, the NB becomes a Poisson distribution. In the presence of overdispersion, the NB count model may be a better fit. After confirming overdispersion in the PRM, the alternative NB model is estimated through the likelihood criterion in Eq. (20), which should also be iterated to a maximum value, that is,

$$\mathcal{L}(\Phi, \beta) = \sum_{i=1}^{n} \left[\left(\sum_{j=0}^{y_i-1} \log(\Phi + j)\right) - \log(y_i!) + \Phi\log(\Phi) + y_i(x_i'\beta) - (\Phi + y_i)\log\left(\Phi + \exp\left\{x_i'\beta\right\}\right)\right] = \max. \tag{21}$$

In our case the iterative algorithm used was IRWLS.

Zero inflated models

In the presence of excess zeros, count data are often not efficiently handled by the standard Poisson and NB model [9]. The Zero inflated models are statistical models that analyse count data exhibiting overdispersion and an excess of zero counts. We therefore further propose the ZINB and ZIP regression models which are often used when the response count variable exhibits over dispersion and has excess zeros. The zero inflated models assume a two part model processes which are estimated from the mixture of binary distribution that is degenerate at 0 (structural zeros) and Poisson gamma distribution for ZINB or Poisson distribution for positive counts including sampling zeros [7]. We suppose that

$$y_i = \begin{cases} 0 & \text{with probability } p \\ \text{Poisson or NegBin} & \text{with probability } 1 - p \end{cases}, \tag{22}$$

where p is the logit link modelled function of predictors driving the excess zeros. Employing the ZIP and ZINB models with the distinctive ability to account for structural zeros (which represent a scenario where the absence of involved axillary lymph nodes is inherent) and sampling zeros which, despite the potential for involvement, no nodes were observed as involved, was imperative as both were critical components to consider in breast cancer patients.

Zero Inflated Poisson (ZIP)

: In the ZIP model, two processes are assumed to generate the zero values. The first process is due to the structural zeros using the binomial distribution and the other process is due to the sampling zeros using the Poisson distribution. Structural zeros are generated using the logit link function, capturing a probability p_i of a structural zero, whilst the modeling of the number of lymph nodes involved with positive counts is done using the log-linear equation with a log-link function for μ_i in the Poisson model. Assuming the involved nodal count data follows a ZIP distribution we have the model $Y_i \sim ZIP(\mu_i, p_i)$. The Poisson general linear model is given as:

$$f(y_i; \mu_i \mid y_i \geq 0) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}. \tag{23}$$

The probability that the nodal count process gives a zero is given as:

$$P(y_i = 0) = p_i + (1 - p_i)e^{\mu_i}. \tag{24}$$

The probability of observing a non-zero count is given by:

$$P(Y_i = y_i | y_i > 0) = (1 - p_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}. \tag{25}$$

The probability mass function of the ZIP is formulated in Eq. (26) assuming a combination of the two process zero generation, one for structural zeros and the other for sampling zeros mentioned earlier:

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)f(y_i) & \text{if } y_i = 0 \\ (1 - p_i)f(y_i) & \text{if } y_i = 1, 2, 3, \ldots \end{cases}, \tag{26}$$

where $f(y_i)$ is the probability function of the Poisson distribution. The structural zeros are generated using the logit link function for p_i expressed as:

$$\text{logit}(p_i) = \tau_i, \qquad p_i = \frac{\lambda_i}{1+\lambda_i}, \tag{27}$$

with

$$\lambda_i = \exp \tau_0 + \tau_1 z_{i1} + \cdots + \tau_m z_{im}. \tag{28}$$

The $z^{'}s$ in Eq.(28) are the set of m latent covariates in the logistic component of the model where z_i is the i^{th} row of the data matrix of the logit model generating the structural zeros, which play a pivotal role in providing insights into the likelihood of observing zero counts. These covariates influence the occurrence of structural zeros, with the regression coefficients used to establish the probability structured zeros, p_i. On the other hand, the sampling zeros and positive count data is modelled using a log-linear equation with a log-link function for μ_i, expressed as:

$$log(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad \Rightarrow \quad \mu_i = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p). \tag{29}$$

Based on Eq. (26) we can formulate the joint likelihood function as:

$$l = \sum_{i=1}^{n} \left(y_i \log(p_i) + \sum_{i=1}^{n} (1-y_i) \log(1-p_i) \right), \tag{30}$$

which results in the following log likelihood:

$$\begin{aligned} L(\beta, \tau; Y_i) = & \sum_{i:y_i=0} \left[\log \left(\exp(z_i'\tau) + \exp\left(-x_i'\beta\right) \right) \right] \\ & + \sum_{i:y_i \geq 0} \left[y_i x_i'\beta - \exp(x_i'\beta) - \log(y_i!) \right] \\ & - \sum_{i=1}^{n} \log\left(1 + \exp(z_i'\tau)\right). \end{aligned} \tag{31}$$

The estimates of the unknown regression coefficients $\beta_0, \beta_1, ..., \beta_p, \tau_0, \tau_1, ..., \tau_m$ of the model are obtained using the maximum likelihood estimation given that y_i are independent observations with corresponding X_i values of the covariates. Using Eq. (27) and Eq. (29) we may define the occurrence of structural zeros being influenced by the covariates z_i's and the occurrence of a particular count y_i including sampling zeros is impacted by x_i's covariates. Covariates can be associated with the probability of a structured zero, p_i, as well as the mean function μ_i of the count model. p_i is modeled with a logistic regression and μ_i is modelled with a log-linear equation. The coefficients for the zero inflated components give the log odds of observing a zero count instead of a positive count with a positive coefficient showing that the covariate increases the odds of observing a zero count. A negative coefficient shows that the covariate decreases the odds of obtaining a zero count. In the Poisson or NB model component, the coefficients represent the expected change in the log of the outcome involved nodal count variable for one unit increase in the covariate, where a positive coefficient indicates that the covariate is associated with an increase in the expected involved nodal count. A negative coefficient shows that the covariate is associated with a decrease in the expected involved nodal count. We used a widely used Broyden, Fletcher, Goldfarb, and Shano (BFGS) [37], a quasi-Newton optimization method to maximize the likelihood functions to get the parameter estimates, as this technique has fewer constraints compared to other methods in a small sample. Quasi-Newton method, which are algorithms which modify Newton's method are fast and accurate approximation techniques.

Zero-inflated negative binomial (ZINB)

: The outcome Y_i for individual i is assumed to follow a negative binomial distribution for positive counts in the ZINB, thus $f(y_i)$ in Eq.(26) is the negative binomial distribution earlier discussed, given in Eq. (15). In a similar approach to the ZIP model, the log link (μ_i) and logit link function p_i is then used as represented in Eq. (27) and Eq. (29).

Using Eq. (26), we also estimate the unknown coefficients $\phi, \beta_0, \beta_1, ..., \beta_p, \tau_0, \tau_1, ..., \tau_m$ of the ZINB model with the maximum values for the log likelihood function in Eq. (32) of the ZINB model. In this case the log-likelihood function is given by:

$$\begin{aligned} \mathcal{L}(\Phi, \beta, \tau; Y_i) = & \sum_{i:y_i=0} \ln \left[p_i + (1 + \Phi^{-1}\mu_i)^{-\Phi} \right] \\ & + \sum_{i:y_i=1,2,3,\ldots} \sum_{j=0}^{y_i-1} \ln(j+\Phi) \\ & + \sum_{i:y_i=1,2,3,\ldots} \Big\{ -\ln(y_i!) - (y_i+\Phi)\ln(1+\Phi^{-1}\mu_i) \\ & + y_i \ln(\Phi^{-1}) + y_i \ln(\mu_i) \Big\} \\ & + \sum_{i=1}^{n} \ln(1+p_i). \end{aligned} \tag{32}$$

Hurdle models

In contrast to zero-inflated models, hurdle models which assume all '0' data are structural, are viewed as a two-component mixture model consisting of zero's and the positive (i.e. non-zero) following either truncated Poisson or truncated NB distribution. The first is a binary model to estimate binary process of zero counts versus positive counts. The second is a zero-truncated Poisson or NB model to estimate over-dispersed positive counts. The hurdle model allows different parameters for specifying zero and the truncated part. The HP model can be written as:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0 \\ (1-p_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!(1-e^{-\mu_i})} & \text{if } y_i = 1, 2, 3, \ldots \end{cases}. \tag{33}$$

The HNB model is represented as:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0 \\ \frac{(1-p_i)}{1-\left(\frac{Y}{\mu_i+Y}\right)^Y} \cdot \frac{\Gamma(y_i+Y)}{\Gamma(Y)\,y_i!} \cdot \left(\frac{\mu_i}{\mu_i+Y}\right)^{y_i} \cdot \left(\frac{Y}{\mu_i+Y}\right)^Y & \text{if } y_i = 1, 2, \ldots, \end{cases} \tag{34}$$

where Y is the dispersion parameter. Hence the hurdle model can be written as:

$$logit(p_i) = Z_i^T \beta, \tag{35}$$

$$\log(\mu_i) = X_i^T \alpha. \tag{36}$$

2.3. PERFORMANCE EVALUATION

In a study that applies several count model frameworks to a dataset, the performance of the models can be compared using different goodness of fit measures as vastly described in the literature. This gives guidance to the most appropriate model which can be implemented in line with the specific nature of the dataset. Some of the popularly used measures include the -2 log likelihood (also called the deviance) , Akaike Information Criterion (AIC) as formulated by Akaike [38], Bayesian information Criterion (BIC) first designed by Schwarz [39] and Akaike Information Criterion with small sample correction (AICc) which we also applied in this study. The deviance statistic compares the difference in probability between the predicted and actual involved axillary lymph nodes for each y_i and sums the differences together to provide a measure of the total error on the count models. In the zero inflated models, the -2 log likelihood, a measure of the total deviance, is the sum of the deviance associated with the involved nodal count model and the deviance associated with the zero inflation model expressed as:

$$-2 \log L = -2 \sum_i \left(y_i \log(\mu_i) - \mu_i - \log(y_i!) \right), \tag{37}$$

where L is the likelihood function of the model, μ_i and y_i are the predicted and the observed terms of the involved nodal count outcome variable for i^{th} observation respectively based on the model.

The AIC is calculated as:

$$AIC(m) = -2[k - \ln L(m)], \tag{38}$$

where k is the number of covariates in the study and $L(m)$ is the log-likelihood estimate for model m. The model with the lowest AIC is considered a better fit.

AICc is defined using the AIC as:

$$AICc(m) = AIC(m) + \frac{2k(k+1)}{n} - k - 1, \tag{39}$$

n is the sample size and $\frac{2k(k+1)}{n} - k - 1$ is the correction term which adjusts the AIC for small sample size bias. Bayesian information Criterion (BIC) is given by:

$$BIC(m) = -2 \log L(m) \log(n) k. \tag{40}$$

2.4. BOOTSTRAP RESAMPLING OF OBSERVATIONS

After getting the count regression performance results, model validation is the next crucial step to assess whether they are likely to hold outside the original data, in this study through a bootstrap resampling of observations [26]. This approach involves collecting the response and explanatory values for each i observation:

$$z_i' = y_i, x_{i1}, x_{i2} \ldots x_{iq}. \tag{41}$$

Repeating the bootstrap sampling process results in many bootstrap samples (r bootstrap samples) that are assumed to represent alternative sample that might have occurred instead of the original sample and these bootstrap samples can be used to estimate properties of estimation process, such as regression coefficient estimates, standard errors and bootstrap confidence intervals. The original n observations can be resampled with replacement, under equal probability $\frac{1}{n}$, yielding bootstrap new sets of r samples of $z_{b1}'^*, \ldots, z_{bn}'^*$, ensuring that each sample is the same size as the original dataset. Some researchers have suggested varying minimum number of iterations to fit the model [27, 40]. Here, 5000 bootstrap samples were generated from the original dataset. For each bootstrap sample, the model is fitted to predict the number of involved axillary lymph nodes. The coefficients from each fitted model is extracted and stored. This process is repeated for all 5000 bootstrap samples, resulting in a distribution of coefficient estimates for each predictor. This leads to the estimation process of the regression coefficient estimates, standard errors and bootstrap confidence intervals. For each replicate $b_1, b_2, \ldots b_r$, the response mean estimate $\hat{\mu}_b^*$ and the regression coefficient estimates

$$\mathbf{b}_{bj}^* = [b_{b0}^*, \ldots, b_{bq}^*]', \tag{42}$$

can be computed.

The bootstrap averages across r bootstrap replicates for the response mean

$$\hat{\mu}_b^* = \frac{1}{r} \sum_{b=1}^{r} \hat{\mu}_{bi}^*, \tag{43}$$

and the regression coefficients are given by:

$$\mathbf{b}_b^* = \frac{1}{r} \sum_{b=1}^{r} \mathbf{b}_{bj}^*. \tag{44}$$

Their standard errors can be approximated as:

$$SE(\hat{\mu}_b^*) = \sqrt{\frac{1}{r-1} \sum_{b=1}^{r} (\hat{\mu}_{bi}^* - \hat{\mu}_b^*)^2}, \tag{45}$$

for the bootstrap mean and

$$SE(\mathbf{b}_b^*) = \sqrt{\frac{1}{r-1} \sum_{b=1}^{r} (\mathbf{b}_{bj}^* - \mathbf{b}_b^*)(\mathbf{b}_{bj}^* - \mathbf{b}_b^*)'}, \tag{46}$$

for the regression coefficients.

The 95% percentile bootstrap confidence interval, a non-parametric approach, was constructed from $\mathbf{b}_{b(lower)}^*$ <b< $\mathbf{b}_{b(upper)}^*$ where $\mathbf{b}_{b(lower)}^* = \mathbf{b}_b^*$ at $0.025r$ i.e the value at the 2.5th percentile and $\mathbf{b}_{b(upper)}^* = \mathbf{b}_b^*$ at $0.975r$ i.e the value at the 97.5th percentile of r ordered bootstrap replicates [41, 42].

3. RESULTS

The analysis is based on 136 breast cancer patients eligible for this study. The plot in Figure 2 shows the involved axillary nodal count data distribution in breast cancer patients in this study. Zero counts accounted for 31.6% of the observations. The spike at the zero counts is evidence that we have zero inflation in our data.

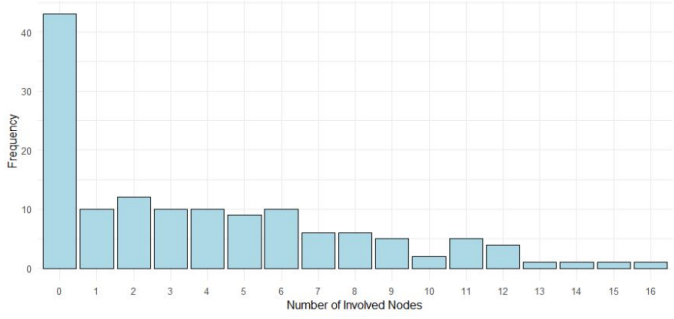


Figure 2. Frequency plot for Axillary Node Counts in Breast Cancer. The horizontal Axis shows the observed axillary nodal count and the vertical axis shows the frequency of observation for each count. This plot highlights the high prevalence of zero counts (patients with no nodal involvement) and a right-skewed distribution, reflecting overdispersion and zero inflation in the data.

Table 1. Measure of over-dispersion.

Model	Poisson	NB	ZIP	ZINB	HP	HNB
Pearson χ^2 value/ DF	4.131	0.7378	1.6696	1.0780	1.5627	1.0010

We first applied a Poisson Regression model to our dataset. The results shown in Table 1 indicated the extent of over-dispersion in the Poisson Regression model with a high value of the dispersion parameter of 4.131 which could not be ignored. We then applied the NB GLM to correct for overdispersion. It is also vital to interpret the dispersion parameters for the NB and Poisson component of the ZIP and ZINB models which represent the overdispersion in the nodal count data. The ratio of the Pearson Chi-Square statistics is just under 1 for the NB (0.7378) and just 1 for the ZINB models (1.078) which is not much of a violation to the dispersion parameter thus no major impact on the results of the zero inflated model, an indication that the ZINB model was able to correct for overdispersion. Similarly, for the HNB the dispersion parameter is approximately 1 (1.010).

Table 2 and Table 3 show the regression coefficients corresponding to covariates used and their corresponding p values. It is noted that in the PRM, the log-linear coefficients show an increase in tumor size ($p = 0.05$) and having tumor grade 2 ($p = 0.09$) was associated with a higher risk of having more involved lymph nodes compared to small size tumors and grade 1 patients respectively (Table 2). For each additional year, the expected nodal count decreased by 46% suggesting older patients tend to have a lower expected involved nodal count compared to younger patients ($p = 0.007$). Being HER2 positive was also a significantly associated with reduced nodal count, having the regression coefficient of -0.255 which equates to an exponentiated coefficient of 0.775. Hence, there is a 22.5% decline in the rate of nodal counts involved compared to patients with a negative status ($p = 0.015$). ER positive and PR positive status all had positive effects in the PRM. The estimate effects for most covariates in the NB were slightly higher compared to those from PRM. In the zero-inflated and hurdle models, there are two sets of components for each model, first showing probability the particular covariate affected the "structural" zeros and the second showing the model for the "sampling" zeros or positive counts for HNB.

Table 2. Output results for Poisson and negative Binomial models for number of involved axillary lymph nodes.

Poisson model				
Variable	Estimate	SE	Z	P(>|z|)
(Intercept)	1.494	0.150	9.946	< 2e-16
Age	-0.616	0.229	-2.688	0.007
Tumor Size	0.482	0.247	1.955	0.051
Tumor Grade (1)	0.161	0.114	1.411	0.158
Tumor Grade (2)	0.149	0.083	1.702	0.089
ER Positive	0.047	0.137	-0.345	0.730
PR Positive	0.196	0.125	1.541	0.123
HER2 Positive	-0.255	0.105	-2.440	0.015
Negative Binomial model				
Variable	Estimate	SE	Z	P(>|z|)
(Intercept)	1.690	0.371	4.561	5.09e-06
Age	-0.942	0.546	-1.724	0.085
Tumor Size	0.396	0.633	0.627	0.531
Tumor Grade (1)	0.163	0.279	0.584	0.559
Tumor Grade (2)	0.159	0.201	0.790	0.430
ER Positive	-0.132	0.328	-0.403	0.687
PR Positive	0.221	0.299	0.737	0.461
HER2 Positive	-0.256	0.246	-1.041	0.298

The count component estimates the expected number of involved axillary lymph nodes. The zero-component estimates the probability of having no involved nodes. Across all the models, age is consistently significant in the zero-component part (Table 3). However, the effect is negative in the HNB. The NB fails to detect a significant effect of age suggesting that considering all zeros as part of the count process may bias the model against detecting an effect of age in this sample. This may be an indication of the presence of structural zeros which the NB does not account for. In the count part of the Poisson models (Poisson, ZIP and HP), an increase in tumor size had a significant multiplicative effect of 1.5 on expected nodal count (p=0.05). The ZIP and HP models' count component revealed a significant positive association for tumor grade with a multiplicative effect of 1.3 for those with grade 2 compared to those with grade 1 tumors. Overall, the Poisson models yield smaller effects compared to the corresponding NB models, possibly suggesting that failure to account for overdispersion by the Poisson models leads to under-estimation of the parameter estimates. It is also noted that the coefficients of zero components are almost similar in HP and HNB (Table 3). This is because the zero part is estimated completely separate from the truncated part.

4. MODEL DIAGNOSTICS

We assessed the performance of our six models using information criteria AIC, AICc and BIC as shown in Table 4 to determine the model with the best fit. The deviance, a measure of goodness of fit based on the -2 log likelihood was also used to evaluate model performance. The NB model and its variations (zero-inflated and hurdle) consistently showed lower deviance, AIC, AICs and BIC compared to the Poisson models. Notably, Poisson model exhibited higher deviance values, indicating a relatively poor fit. Among all the models, the ZINB model had the lowest AIC and BIC suggesting it might be the best fitting

Table 3. Output results for Zero-inflated and hurdle models.

Log Link (count)				
Variable	ZIP	ZINB	HP	HNB
Intercept	1.452	1.383	1.454	1.407
SE	0.159	0.278	0.158	0.272
p-value	<2e-16	6.19e-07	<2e-16	2.24e-07
Age (years)				
Estimate	0.222	0.253	0.211	0.180
SE	0.252	0.438	0.250	0.424
p-value	0.378	0.563	0.399	0.672
Tumor Size (cm)				
Estimate	0.444	0.489	0.444	0.488
SE	0.235	0.405	0.236	0.405
p-value	0.059	0.227	0.059	0.228
Tumor Grade 2				
Estimate	0.248	0.258	0.248	0.264
SE	0.114	0.185	0.115	0.185
p-value	0.030	0.162	0.030	0.154
Tumor Grade 3				
Estimate	0.085	0.064	0.085	0.067
SE	0.085	0.137	0.085	0.136
p-value	0.312	0.647	0.320	0.623
ER Positive				
Estimate	-0.028	-0.081	-0.019	-0.044
SE	0.149	0.256	0.148	0.247
p-value	0.852	0.751	0.897	0.857
PR Positive				
Estimate	0.239	0.307	0.233	0.273
SE	0.134	0.234	0.133	0.226
p-value	0.074	0.188	0.079	0.228
HER2 Positive				
Estimate	-0.175	-0.185	-0.170	-0.176
SE	0.109	0.180	0.109	0.176
p-value	0.113	0.304	0.177	0.317
Zero Count				
Intercept	-2.446	-2.976	2.368	2.368
SE	0.759	0.980	0.768	0.737
p-value	0.001	0.002	0.000	0.001
Age (years)				
Estimate	2.837	3.439	2.272	-2.728
SE	1.042	1.311	0.006	1.006
p-value	0.006	0.009	0.009	0.007
Tumor Size (cm)				
Estimate	0.099	0.257	-0.065	-0.065
SE	1.181	1.396	1.169	1.169
p-value	0.933	0.843	0.960	0.956
Tumor Grade 2				
Estimate	0.416	0.588	-0.379	-0.379
SE	0.564	0.686	0.550	0.550
p-value	0.465	0.416	0.490	0.490
Tumor Grade 3				
Estimate	-0.208	-0.239	0.215	0.215
SE	0.386	0.460	0.377	0.377
p-value	0.590	0.602	0.569	0.569
ER Positive				
Estimate	0.183	0.064	-0.219	-0.219
SE	0.619	0.770	0.600	0.600
p-value	0.768	0.934	0.715	0.715
PR Positive				
Estimate	0.118	0.324	-0.065	-0.065
SE	0.563	0.713	-0.120	0.542
p-value	0.835	0.649	0.904	0.905
HER2 Positive				
Estimate	0.245	0.210	-0.269	-0.269
SE	0.444	0.515	0.434	0.434
p-value	0.581	0.684	0.535	0.535

model. Given the small differences between the ZINB and HNB models and considering the goal of achieving stable and reliable estimates through bootstrapping, both models were selected for further analysis.

Table 4. Regression count model performance comparisons.

Distribution	-2 Log Likelihood	AIC	AICC	BIC
Poisson	898.21	915.26	916.39	938.55
NB	662.50	680.50	681.93	706.72
ZIP	674.20	706.15	710.72	752.75
ZINB	631.90	665.92	671.11	715.44
HNP	674.20	706.23	710.81	752.83
HNB	632.10	666.45	671.64	715.97

Table 5. Bootstrap results using HNB model.

Negative Binomial log link						
Variable	Original Est.	Bootstrap Est.	Bootstrap S.E	Bias	C.I Original Est.	C.I Bootstrap Est.
(Intercept)	1.407	1.381	0.318	-0.021	(0.874, 1.940)	(0.765, 2.024)
Age	0.180	0.193	0.460	0.013	(-0.653, 1.012)	(-0.743, 1.073)
Tumor Size	0.488	0.438	0.432	-0.051	(-0.309, 1.285)	(-0.278, 1.428)
Tumor Grade 2	0.264	0.280	0.197	0.016	(-0.103, 0.630)	(-0.049, 0.720)
Tumor Grade 3	0.067	0.065	0.146	-0.001	(-0.200, 0.333)	(-0.253, 0.324)
ER Positive	-0.044	-0.053	0.292	-0.008	(-0.530, 0.441)	(-0.596, 0.567)
PR Positive	0.272	0.293	0.237	0.020	(-0.173, 0.719)	(-0.197, 0.739)
HER2 Positive	-0.176	-0.168	0.197	0.008	(-0.521, 0.169)	(-0.576, 0.201)
Binomial Logit link						
(Intercept)	2.367	2.788	1.411	0.420	(0.914, 3.822)	(0.590, 4.260)
Age	-2.727	-2.951	1.154	-0.223	(-4.707, -0.748)	(-4.816, -0.351)
Tumor Size	-0.064	-0.112	1.218	-0.047	(2.356, 2.227)	(-2.479, 2.320)
Tumor Grade 2	-0.379	-0.925	2.459	-0.545	(-1.463, 0.705)	(-1.675, 1.122)
Tumor Grade 3	0.214	0.527	1.438	0.313	(-0.532, 0.696)	(-0.720, 1.252)
ER Positive	-0.219	-0.202	0.710	0.015	(-1.394, 0.957)	(-1.644, 1.192)
PR Positive	-0.065	-0.092	0.635	-0.027	(-1.126, 0.997)	(-1.395, 1.097)
HER2 Positive	-0.269	-0.278	0.516	-0.009	(-1.126, 0.588)	(-1.282, 0.759)

At this point, our focus is on finding the best fitting model to our data between the ZINB and HNB models. Bootstrapping was performed using the ZINB and HNB since both had almost similar metrics (Table 4). The HNB model exhibited stability in the bootstrap resampling analysis compared to the ZINB, and we present results only for the HNB analysis in Table 5. The bootstrap estimates, standard error, biases and confidence intervals are presented. The bootstrap estimates for HNB model were consistent with the original estimates, suggesting stable and reliable results in both components of the model. The bias for each coefficient also shows relatively small values, indicating the bootstrap estimates are close with the original estimates. The confidence intervals of a bootstrap distribution were calculated for each coefficient. These intervals are considered robust as they adjust for both bias and skewness in the distribution of the bootstrap estimates. This result can confirm the usefulness of the bootstrap regression coefficients.

Since the HNB model demonstrated stability, results for this study are interpreted from the original model. In the count part, age has a positive effect of 0.180 indicating that an increase in age might slightly increase the number of involved nodes, although not statistically significant. Tumor size shows a positive association, with a unit increase in size resulting in a 63% increase in the expected nodal count (exp (0.488) =1.63). Higher tumor grades (2 and 3) and PR positive status have positive effects on involved axillary nodes in the count component of the HNB model. In the zero-part, the coefficient for age of -2.279 indicates that each additional unit increase in age decreases the odds of having a zero count (no involved lymph nodes) by 2.279 times suggesting that older patients are less likely to have zero counts compared to younger patients. This result is statistically significant

with a p-value of 0.007. Tumor size, moderately defined tumors, ER and HER2neu positive status also have negative effects abeit non-significance. Poorly differentiated tumor a positive effect on nodal count. The HNB model based on parameter estimation is:

$$\begin{aligned} &logit(p_1) = 2.367 - 2.727 * Age - 0.064 * tumor.size \\ &- 0.379 * tumor.grade2 + 0.214 * tumor.grade3 - 0.219 * ER \\ &- 0.065 * PR - 0.269 * HER2 \\ &\log(\mu_i) = 1.407 - 0.180 * Age + 0.488 * tumor.size \\ &+ 0.264 * tumor.grade2 + 0.067 * tumor.grade3 \\ &- 0.044 * ER + 0.272 * PR - 0.176 * HER2. \end{aligned} \quad (47)$$

5. DISCUSSION

Involved axillary lymph nodes are among the most important prognostic factors in women diagnosed with breast cancer [43]. Understanding the factors associated with the number of nodes involved is essential for directing treatment strategies, as these factors play a pivotal role in predicting outcomes and guiding therapeutic interventions. In this study, we compared various count regression models to determine the best fitting model for analyzing the count outcome variable, involved axillary lymph nodes. To ensure the stability and reliability of our parameter estimates, we integrated bootstrap resampling techniques. This methodological approach enhances the validity of our findings, ensuring that our conclusions are robust and less affected by sample variability.

We considered six different models involving either the Poisson or NB distributions for analyzing breast cancer nodal count outcome data. The performance and stability of model coefficients were considered to determine the most appropriate model for our data. A single parameter distribution Poisson model was less suitable in our study data compared to two parameter distributions. The basic PRM yielded undesirable results due to overdispersion prompting a comparison with the alternative NB, ZINB and ZIP models. Although the NB evidently corrected the large over dispersion in the nodal count data compared to PRM, it remained an unsuitable choice due to challenges posed by the zero inflated data. Inspection of the observed data as well as fit statistics suggested that the distribution of the nodal count data was both over-dispersed and zero-inflated. We then fit the zero-inflated and hurdle models and the results showed that models using NB distributions fit better than the Poisson models. The Poisson models consistently produced smaller estimates compared to NB models. These discrepancies can be attributed to the underlying assumptions of these models. Taken together, these suggest the importance of accounting for both over-dispersion and zero-inflation in modelling count data also illustrating the risk of falsely identifying significant effects of variables if the model chosen does not model the distribution of the data correctly.

The two-parameter distribution NB when integrated into a two-part hurdle (HNB) and ZINB provided the best fit as evidenced by having the lowest AIC, BIC and AICc. This dual structure is particularly useful in clinical settings where a significant portion of portion may have no lymph node involvement, yet amongst those with involvement, the count can vary widely. Our findings are similar to the results presented by Liquat *et al.* [16], Pavlicova *et al.* [20] and Andika *et al.* [44] where the zero-augmented NB models performed better than zero-augmented Poisson models. In contrast, a study by Fernandez *et al.* [45] found that the PRM had the smallest AIC and BIC. However, in their study, the simulated data that produced such results had very low overdispersion (0.01), whereas in our study, the dispersion parameter was 4.079.

The bootstrap resampling technique was applied to the best fitting models HNB and ZINB models to evaluate the stability of parameter estimates and select the most appropriate model for the data. The bootstrap results for the ZINB showed unreasonable estimates with instability, very high standard errors and substantial bias particularly in the zero-component. In contrast, the HNB model produced bootstrap estimates that closely matched the original estimates, with minimal bias. The bootstrap confidence intervals in the count component, and most in the zero-component were almost similar to the original confidence intervals, indicating stability of the HNB model. Moreover, the information criteria metrics suggested that relying on single-time modeling could lead to incorrect model selection between HNB and ZINB. Both models may fail to detect under- or overestimation when validated only once. Bootstrap resampling validated the reliability, stability, and consistency of the HNB model, underscoring its robustness in analyzing factors associated with axillary lymph node involvement. This finding highlights the importance of selecting an appropriate model for count outcome data to avoid biased parameter estimates, which can result in either underestimation or overestimation of effects. By integrating bootstrap resampling with advanced count regression models, this study addressed critical challenges in analyzing small, over-dispersed datasets. Traditional applications of zero-inflated and hurdle models often rely on large, balanced datasets, which may not reflect the realities of resource-limited settings like Zimbabwe. Incorporating bootstrap resampling ensures robust parameter estimates and reliable inferences, even in the presence of data limitations, thereby advancing statistical modeling in resource-constrained contexts.

Given its superior performance in stability, reliability, and consistency during bootstrap resampling, the HNB model was selected as the basis for interpreting the relationships between covariates and the count of involved lymph nodes in breast cancer patients. Bootstrap resampling further validated the precision of these estimates, underscoring their robustness. Age was a significant predictor in the zero-inflation component of the HNB model, indicating that older patients are less likely to have axillary lymph nodes affected. This significant negative association suggests that older patients may develop breast cancers that are less likely to spread to the lymph nodes, potentially due to slower tumor progression or other age-related biological factors. These findings align with existing literature emphasizing the role of age-related changes in tumor biology on patterns of disease progression [46–50]. The robustness of this association was supported by bootstrap resampling, which demonstrated narrower confidence intervals for the age coefficient and consistent exclusion of null values, confirming that this relationship is not due to random variability. Interestingly, this study found no association between age and positive nodal counts, a finding that contrasts with studies by Sandoughdaran *et al.* [51] and Greer *et al.* [52], which reported younger age associated with aggressive cancers

and a higher risk of lymph node invasion. These discrepancies may reflect differences in study populations, tumor subtypes, or the statistical methods used. For instance, Sandoughdaran *et al.* [51] and Greer *et al.* [52] focused on larger cohorts with distinct demographic characteristics, which may have influenced their results. Further research is needed to clarify these differences and explore the mechanisms underlying age-related variations in nodal involvement. These findings highlight the importance of tailored screening and early detection programs. Even when the likelihood of nodal involvement is low for older patients, age-related factors influencing breast cancer progression warrant targeted clinical strategies to ensure timely diagnosis and intervention.

While the HNB model provided the best fit for the data, it is noteworthy that tumor grade was significant, and tumor size was marginally significant in the Zero-Inflated Poisson (ZIP) and Hurdle Poisson (HP) models but not in the HNB model. These discrepancies likely reflect differences in how the models handle over-dispersion and excess zeros. Unlike ZIP and HP, the HNB model accounts more effectively for over-dispersion and zero inflation, which may explain the lack of significance for tumor size. This underscores the importance of selecting appropriate models for analyzing count data in medical research, as model assumptions can influence the interpretation of results. Although tumor size was not statistically significant in the HNB model, the positive estimate in the count component still indicates an increased risk of lymph node involvement with larger tumor size. This aligns with findings from previous studies [53, 54]. However, the consistent lack of significance in the zero component across all models suggests that tumor size primarily affects the extent of positive nodal involvement rather than the probability of zero involvement.

Contrary to our findings, Swain *et al.* [14] reported tumor size as a significant predictor of lymph node involvement. Similarly, Shima *et al.* [24] found significant associations between metastasis status, HER2-positive status, higher tumor grade, and increased risk of involved lymph nodes using the ZINB model in an Iranian population. These discrepancies highlight the importance of considering region-specific models, as variations in sample size, demographic, and clinical characteristics across studies can lead to differences in significance levels. Our study population, consisting of breast cancer patients from a resource-limited setting in Zimbabwe, likely has unique demographic and clinical profiles compared to other populations. Additionally, inherent delays in diagnostic processes within this setting may have influenced our findings, further underscoring the need for context-specific analyses.

Unlike studies that relied on binary classification to assess nodal status, such as that by Elleson *et al.* [55], Jiang *et al.* [56] and Zhang *et al.* [29], which may oversimplify the heterogeneity in medical data, this study utilized zero-inflated and hurdle count models. These models offered a more nuanced perspective by not only analyzing the count of involved lymph nodes but also explaining the probability of zero nodal involvement. The zero-inflation component of the HNB model provided critical insights into factors influencing the likelihood of zero nodal involvement, a perspective often overlooked in simpler analyses. These flexible models can support clinicians in making informed, data-driven decisions about targeted treatments for breast cancer patients, particularly in resource-constrained settings.

However, our study had some limitations. We used a cross sectional study dataset with limited data on vital characteristics and the use of secondary data introduced the possibility of human capturing errors. The hospital where the data was extracted is a national referral hospital in Zimbabwe. This means we were dealing with a select population hence the results in our analysis may only be applicable to referrals and not for the general population.

6. CONCLUSION

Traditional models like Poisson and NB were initially considered but proved inadequate due to overdispersion and excess zero counts. Zero-inflated and hurdle models, specifically designed to handle these complexities were explored, with the HNB model emerging as the most reliable based on both model fit and bootstrap resampling, which addressed instability observed in the ZINB model. This study highlights the importance of selecting appropriate models for count data in medical research and demonstrates the utility of bootstrap resampling for enhancing the reliability of estimates in small, variable datasets. By addressing the challenges posed by small sample sizes and complex data structures, this approach contributes to more reliable and actionable insights in medical research.

To translate these findings into practice, we recommend integrating bootstrap resampling into count regression modeling to improve the accuracy of clinical factor estimates. This approach can support treatment planning, policy evaluation, and resource allocation, ultimately improving patient care. Age was identified as a significant predictor of zero nodal involvement, underscoring the need for health policies that prioritize early detection programs, particularly for younger patients, alongside improved access to screening and diagnostic services. Accurate tumor characterization remains a critical determinant of disease progression. Patients with higher tumor sizes may benefit from personalized treatment plans and close follow-up to optimize outcomes. Community education campaigns are vital to raise awareness about breast cancer symptoms, promote regular screening, and emphasize the importance of timely treatment. Collaboration between healthcare providers, researchers, and policymakers is essential to address breast cancer challenges. Further studies should validate these findings and explore additional factors influencing nodal involvement. By implementing these recommendations, healthcare systems in resource-limited settings can improve breast cancer outcomes and ensure equitable care for affected patients. This study provides a foundation for future research and policy initiatives aimed at reducing the burden of breast cancer and enhancing the quality of care for patients globally.

ACKNOWLEDGMENT

The authors would like to thank Medical Research Council of Zimbabwe, Parirenyatwa Group of Hospitals (Zimbabwe) and their personnel for the support in providing data and relevant information for the successful execution of this research.

DATA AVAILABILITY STATEMENT

All the data required for the publication has been included within the research article. These are in the form of figures and tables in the research paper. The original dataset supporting the conclusions of this article is not publicly available due to confidentiality but is available from the corresponding author on reasonable request.

References

[1] World Health Organization, ''Zimbabwe burden of cancer: cancer country profile 202'', 2020. [Online]. Available: https://cdn.who.int/media/docs/default-source/country-profiles/cancer/zwe-2020.pdf.

[2] International Agency for Research on Cancer, ''World cancer day: breast cancer overtakes lung cancer in terms of number of new cancer cases worldwide'', 2021, [Press Release]. [Online]. https://www.iarc.who.int/wp-content/uploads/2021/02/pr294_E.pdf.

[3] M. B. Amin, F. L. Greene, S. B. Edge, C. C. Compton, J. E. Gershenwald, R. K. Brookland, L. Meyer, D. M. Gress, D. R. Byrd & D. P. Winchester, ''The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more ''personalized'' approach to cancer staging'', CA: A Cancer Journal for Clinicians **67** (2017) 93. https://doi.org/10.3322/caac.21388.

[4] F. Peintinger, R. Reitsamer., M. Smidt, T. Kuhn & C. Liedtke, ''Lymph nodes in breast cancer - what can we learn from translational research'', Breast Care **13** (2018) 342. https://doi.org/10.1159/000492435.

[5] Y. Zou, X. Hu & X. Deng, ''Distant lymph node metastases from breast cancer-is it time to review TNM cancer staging?'', JAMA Network Open **4** (2021) e212026. https://doi.org/10.1001/jamanetworkopen.2021.2026.

[6] J. M. Hilbe, *Negative binomial regression*, Cambridge University Press, Cambridge, UK, 2011. https://doi.org/10.1017/CBO9780511973420.

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

[8] W. H. Greene, ''Accounting for excess zeros and sample selection in poisson and negative binomial regression models'', NYU Working Paper No. EC-94-10, 1994. [Online]. https://ssrn.com/abstract=1293115.

[9] S. G. Heeringa, B. T. West & P. A. Berglund, *Applied survey data analysis*, Chapman & Hall/CRC, CRC Press, 2017, pp. 1–590. https://doi.org/10.1201/9781420080674

[10] D. C. Heilbron, ''Zero-altered and other regression models for count data with added zeros'', Biometrical Journal **36** (1994) 531. https://doi.org/10.1002/bimj.4710360505.

[11] J. Mullahy, ''Specification and testing of some modified count data models'', Journal of Econometrics **33** (1986) 341. https://doi.org/10.1016/0304-4076(86)90002-3.

[12] O. S. Adesina, ''Bayesian multilevel models for count data'', Journal of the Nigerian Society of Physical Sciences **3** (2021) 224. https://doi.org/10.46481/jnsps.2021.168.

[13] O. Maxwell, B. A. Mayowa, I. U. Chinedu & E. Amadi, ''Modelling count data; a generalized linear model framework'', American Journal of Mathematics and Statistics **8** (2018) 179. https://doi.org/10.5923/j.ajms.20180806.03.

[14] P. K. Swain, S. Grover, S. Chakravarty, K. Goel & V. Singh, ''Estimation of number of involved lymph nodes in breast cancer patients using Bayesian regression approach'', J. Stat. Appl. Probab. Lett. **4** (2017) 17. https://doi.org/10.18576/jsapl/040103.

[15] C. X. Feng, ''A comparison of zero-inflated and hurdle models for modeling zero-inflated count data'', Journal of Statistical Distributions and Applications **8** (2021) 8. urlhttps://doi.org/10.1186/s40488-021-00121-4.

[16] M. Liaqat & S. Kamal, ''Zero-inflated and hurdle models with an application to the number of involved axillary lymph nodes in primary breast cancer'', Journal of King Saud University-Science **34** (2022) 101932. https://doi.org/10.1016/j.jksus.2022.101932.

[17] N. S. Abu Bakar, J. Ab Hamid, M. N. S. M. ShaifulJefri, M. N. Sham & J. A. Syakira, ''Count data models for outpatient health services utilisation'', BMC Medical Research Methodology **22** (2022) 261. https://doi.org/10.1186/s12874-022-01733-3.

[18] K. Nketia & D. K. de Souza, ''Using zero-inflated and hurdle regression models to analyze schistosomiasis data of school children in the southern areas of Ghana'', PLOS ONE **19** (2024) e0304681. https://doi.org/10.1371/journal.pone.0304681.

[19] A. A. Yirga, S. F. Melesse, H. G. Mwambi & D. G. Ayele, ''Negative binomial mixed models for analyzing longitudinal CD4 count data'', Scientific Reports **10** (2020) 16742. https://https://doi.org/10.1038/s41598-020-73883-7.

[20] M.-C. Hu, M. Pavlicova & E. V. Nunes, ''Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial'', The American Journal of Drug and Alcohol Abuse **37** (2011) 367. https://doi.org/10.3109/00952990.2011.597280.

[21] S. Sharker, L. Balbuena, G. Marcoux & C. X. Feng, ''Modeling sociodemographic and clinical factors influencing psychiatric inpatient service use: a comparison of models for zero-inflated and overdispersed count data'', BMC Medical Research Methodology **20** (2020) 232. https://doi.org/10.1186/s12874-020-01112-w.

[22] F. Tüzen, S. Erbaş & H. Olmuş, ''A simulation study for count data models under varying degrees of outliers and zeros'', Communications in Statistics-Simulation and Computation **49** (2020) 1078. https://doi.org/10.1080/03610918.2018.1498886.

[23] L. Xu, A. D. Paterson, W. Turpin & W. Xu., ''Assessment and selection of competing models for zero-inflated microbiome data'', PLoS One **10** (2015) e0129606. https://doi.org/10.1371/journal.pone.0129606.

[24] S. Younespour, E. Maraghi, A. Saki Malehi, M. Raissizadeh, M. Seghatoleslami & M. Hosseinzadeh, ''Evaluating related factors to the number of involved lymph nodes in patients with breast cancer using zero-inflated negative binomial regression model'', Journal of Biostatistics and Epidemiology **6** (2021) 259. https://doi.org/10.18502/jbe.v6i4.5679.

[25] N. Asiamah, H. Kofi Mensah & E. Fosu Oteng-Abayie, ''Do larger samples really lead to more precise estimates? A simulation study'', American Journal of Educational Research **5** (2017) 9. https://doi.org/10.12691/education-5-1-2.

[26] B. Efron, *The Jackknife, the bootstrap and other resampling plans*, Society for Industrial and Applied Mathematics, SIAM, Carlifonia, 1982. https://epubs.siam.org/doi/book/10.1137/1.9781611970319.

[27] N. V. Truong, T. Shimizu & T. Kurihara, ''Generating reliable tourist accommodation statistics: bootstrapping regression model for overdispersed long-tailed data'', Journal of Tourism, Heritage & Services Marketing (JTHSM) **6** (2020) 30. https://doi.org/10.5281/zenodo.3837608.

[28] J. Sillabutra, P. Kitidamrongsuk, C. Ujeh, S. Sae-tang & K. Donjdee, ''Bootstrapping with R to make generalized inference for regression model'', Procedia Computer Science **86** (2016) 228. https://pdf.sciencedirectassets.com/.

[29] X. Zhang, J. Lee & W. W. B. Goh, ''An investigation of how normalisation and local modelling techniques confound machine learning performance in a mental health study'', Heliyon **8** (2022) e09502. https://10.1016/j.heliyon.2022.e09502.

[30] A. Fitrianto, ''A study of count regression models for mortality rate'', CAUCHY: Jurnal Matematika Murni dan Aplikasi **7** (2021) 142. https://doi.org/10.18860/ca.v7i1.13642.

[31] M. Devidas & E. O. George, ''Monotonic algorithms for maximum likelihood estimation in generalized linear models'', Sankhyā: The Indian Journal of Statistics, Series B **61** (1999) 382. http://www.jstor.org/stable/25053099.

[32] J. W. Hardin & J. M. Hilbe, *Generalized linear models and extensions*, Stata Press, Arizona, 2007. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.stata-press.com/books/preview/glmext4-preview.pdf.

[33] S. A. Klugman, H. Panjer & G. Willmot, *Loss models: from data to decisions*, 3rd ed., Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, N.J., 1998. https://lccn.loc.gov/2018031122.

[34] J. M. Hilbe, *Modeling count data*, International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_369.

[35] M. B. Morrissey & G. D. Ruxton, ''Revisiting advice on the analysis of count data'', Methods in Ecology and Evolution **11** (2020) 1133. https://doi.org/10.1111/2041-210X.13473.

[36] J. F. Lawless, ''Negative binomial and mixed poisson regression'', The Canadian Journal of Statistics / La Revue Canadienne de Statistique **15** (1987) 209. https://doi.org/10.2307/3314912.

[37] J. Brownlee, ''A gentle introduction to the BFGS optimization algorithm: tutorial on optimization''. Accessed on 19 May 2021. [Online]. https://machinelearningmastery.com/bfgs-optimization-in-python/.

[38] H. Akaike, ''A new look at the statistical model identification'', IEEE Transactions on Automatic Control **19** (1974) 716. https://doi.org/10.1109/TAC.1974.1100705.

[39] G. Schwarz, ''Estimating the dimension of a model'', Annals of Statistics **6** (1978) 461. https://doi.org/10.1214/aos/1176344136.

[40] L. Simar & P. W. Wilson, ''Estimation and inference in two-stage, semi-parametric models of production processes'', Journal of Econometrics **136** (2007) 31. https://doi.org/10.1016/j.jeconom.2005.07.009.

[41] T. J. DiCiccio & B. Efron, ''Bootstrap confidence intervals'', Statistical Science **11** (1996) 189. https://DOI:10.1214/ss/1032280214.

[42] M. R. Chernick, *Bootstrap methods: a guide for practitioners and researchers*, John Wiley & Sons, 2007. https:doi.org//10.1002/9780470192573.

[43] I. Jatoi, S. G. Hilsenbeck, G. M. Clark & C. K. Osborne, ''Significance of axillary lymph node metastasis in primary breast cancer'', Journal of Clinical Oncology **17** (1999) 2334. https://doi.org/10.1200/JCO.1999.17.8.2334.

[44] A. Andika, S. Abdullah & S. Nurrohmah, ''Hurdle negative binomial regression model'', in ICSA-International Conference on Statistics and Analytics, 2019. [Online]. https://doi.org/10.29244/icsa.2019.pp57-68.

[45] G. A. Fernandez & K. P. Vatcheva, ''A comparison of statistical methods for modeling count data with an application to hospital length of stay'', BMC Medical Research Methodology **22** (2022) 211. https://doi.org/10.1186/s12874-022-01685-8.

[46] S. M. Downs-Canner, C. E. Gaber, R. J. Louie, P. D. Strassle, K. K. Gallagher, H. B. Muss & D. W. Ollila, ''Nodal positivity decreases with age in women with early-stage, hormone receptor-positive breast cancer'', Cancer **126** (2020) 1193. https://doi.org/10.1002/cncr.32668.

[47] M. Luo, X. Lin, D. Hao, K. W. Shen, W. Wu, L. Wang, S. Ruan & J. Zhou, ''Incidence and risk factors of lymph node metastasis in breast cancer patients without preoperative chemoradiotherapy and neoadjuvant therapy: analysis of SEER data'', Gland Surgery **12** (2023) 1508. https://doi.org/10.21037/gs-23-258.

[48] H. Wildiers, B. Van Calster, L. V. van de Poll-Franse, W. Hendrickx, J. Røislien, A. Smeets, R. Paridaens, K. Deraedt, K. Leunen & C. Weltens, ''Relationship between age and axillary lymph node involvement in women with breast cancer'', Journal of Clinical Oncology **27** (2009) 2931. https://doi.org/10.1200/JCO.2008.16.7619.

[49] X. Cui, H. Zhu & J. Huang, ''Nomogram for predicting lymph node involvement in triple-negative breast cancer'', Frontiers in Oncology **10** (2020) 608334. https://doi.org/10.3389/fonc.2020.608334.

[50] Z. Lv, W. Zhang, Y. Zhang, G. Zhong, X. Zhang, Q. Yang & Y. Li, ''Metastasis patterns and prognosis of octogenarians with metastatic breast cancer: A large-cohort retrospective study'', PLOS ONE **17** (2022) e0263104. https://doi.org/10.1371/journal.pone.0263104.

[51] S. Sandoughdaran, M. Malekzadeh & M. E. Akbari, ''Frequency and predictors of axillary lymph node metastases in Iranian women with early breast cancer'', Asian Pacific Journal of Cancer Prevention: APJCP **19** (2018) 1617. https://doi.org/10.22034/APJCP.2018.19.6.1617.

[52] L. T. Greer, M. Rosman, W. C. Mylander, W. Liang, R. R. Buras, A. B. Chagpar, M. J. Edwards & L. Tafra, ''A prediction model for the presence of axillary lymph node involvement in women with invasive breast cancer: A focus on older women'', The Breast Journal **20** (2014) 147. https://doi.org/10.1111/tbj.12233.

[53] V. Sopik & S. A. Narod, ''The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer'', Breast Cancer Research and Treatment **170** (2018) 647. https://doi.org/10.1007/s10549-018-4796-9.

[54] S. K. Min, S. K. Lee, J. Woo, S. M. Jung, J. M. Ryu, J. Yu, J. E. Lee, S. W. Kim, B. J. Chae & S. J. Nam, ''Relation between tumor size and lymph node metastasis according to subtypes of breast cancer'', Journal of Breast Cancer **24** (2021) 75. https://doi.org/10.4048/jbc.2021.24.e4.

[55] K. M. Elleson, K. Englander, J. Gallagher, N. Chintapally, W. Sun, J. Whiting, M. Mallory, J. Kiluk, S. Hoover, N. Khakpour & others, ''Factors predictive of positive lymph nodes for breast cancer'', Current Oncology **30** (2023) 10351. https://doi.org/10.3390/curroncol30120754.

[56] C. Jiang, Y. Xiu, K. Qiao, X. Yu, S. Zhang & Y. Huang, ''Prediction of lymph node metastasis in patients with breast invasive micropapillary carcinoma based on machine learning and SHapley Additive exPlanations framework'', Frontiers in Oncology **12** (2022) 981059. https://doi.org/10.3389/fonc.2022.981059.

[57] Y. Zhang, J. Li, Y. Fan, X. Li, J. Qiu, M. Zhu & H. Li, ''Risk factors for axillary lymph node metastases in clinical stage T1-2N0M0 breast cancer patients'', Medicine **98** (2019) e17481. https://doi.org/10.1097/MD.0000000000017481.