Journal of Data Science **12**(2014), 307-323

On Classifying At Risk Latent Zeros Using Zero Inflated Models

Alok Kumar Dwivedi [1*], MB Rao[3], Sada Nand Dwivedi[2],
S.V. S. Deo[2] and Rakesh Shukla[3]

[1]*Texas Tech University Health Sciences Center*
[2] *All India Institute of Medical Sciences*
[3]*University of Cincinnati*

Abstract: Count data often have excess zeros in many clinical studies. These zeros usually represent "disease-free state". Although disease (event) free at the time, some of them might be at a high risk of having the putative outcome while others may be at low or no such risk. We postulate these zeros as a one of the two types, either as 'low risk' or as 'high risk' zeros for the disease process in question. Low risk zeros can arise due to the absence of risk factors for disease initiation/progression and/or due to very early stage of the disease. High risk zeros can arise due to the presence of significant risk factors for disease initiation/ progression or could be, in rare situations, due to misclassification, more specific diagnostic tests, or below the level of detection. We use zero inflated models which allows us to assume that zeros arise from one of the two separate latent processes-one giving low-risk zeros and the other high-risk zeros and subsequently propose a strategy to identify and classify them as such. To illustrate, we use data on the number of involved nodes in breast cancer patients. Of the 1152 patients studied, 38.8% were node- negative (zeros). The model predicted that about a third (11.4%) of negative nodes are "high risk" and the remaining (27.4%) are at "low risk" of nodal positivity. Posterior probability based classification was more appropriate compared to other methods. Our approach indicates that some node negative patients may be re-assessed for their diagnosis about nodal positivity and/or for future clinical management of their disease. The approach developed here is applicable to any scenario where the disease or outcome can be characterized by count-data.

Key words: Count data, Classification, Low-risk zeros, High-risk zeros, Zero inflated model

1. Introduction

Classification of individuals as such who are at high-risk of certain outcome is an important goal in clinical practice and research (Lewis, 2000). Classification of individuals with non-disease/negative outcome into low and high-risk group is equally important for further management of the patients so that individuals, who are event-free but at high risk, can be monitored more closely than those with low risk of the event. We propose an approach to classify patients with negative outcome who are, nevertheless, at high-risk if outcome is measured as ordinal or count data.

Count outcome data often occur in medical research for example, number of days with physical activities, number of adverse cardiac events, number of recurrences, number of attacks, number of seizure in epilepsy, number of hospital admissions, number of alcoholic drinks consumed etc. Data collected on such outcomes often have excess of zeros (negative outcomes, no disease, or no event) (Slymen et al., 2006). Excess zeros in count outcome data may occur due to presence of more subjects with no risk of event of interest. Studies refer to such zeros as structural zeros and zeros that are at risk of event in question are referred to as sampling zeros. Structural zeros are also referred to as true/positive/non-random zeros whereas sampling zeros are also referred to as false/negative/random/chance zeros. It is difficult to differentiate structural and sampling zeros from the observed zeros alone. If we know in advance, some or all-structural zeros then we could eliminate these zeros from the model (Mohri and Brian, 2005). Regardless of the types of zeros, there is a need to account for the data heterogeneity due to excess zeros in drawing appropriate inferences and predictions while modeling count data.

In clinical studies, we often deal with disease problems having count outcomes in which each individual is at risk of disease/outcome of interest. In such scenarios, all zeros are at risk zeros. In such clinical research data situations, and if there are excess zeros, then we can define zeros into two classes (low/no risk or high risk) (Dwivedi et al., 2010). These zeros may arise due to many reasons in a disease process. If zeros arise due to absence of risk factors for disease initiation/progression and/or due to early detection of disease then it may be defined as low/no risk zeros. If zeros arise due to the presence of significant risk factors for disease initiation, progression, due to misclassification, due to more specific/refined diagnostic tests or due to below the level of detection then it can be defined as high risk zeros. High risk zeros may be more likely to develop event of interest in near future. Thus, there is a need to classify these latent zeros as such so that the subjects with high risk zeros can be managed accordingly. Consequently, event of interest may be minimized or prevented among the subjects with high risk zeros.

Count data involving excess zeros is usually described using zero hurdle or inflated count models to account for the variability due to excess zeros. It has been suggested that zero hurdle models are more appropriate in case of one kind of zeros while zero inflated models should be preferred in case of mixtures of zeros i.e., involvement of both types of zeros (Rose et al., 2006). In any situation of excess zeros, we can classify all zeros into two latent groups (no/low-risk and high-risk) and subsequently, we can use Zero Inflated (ZI) models. In ZI models, excess zeros that arise from the non-counting process can be termed as low-risk zeros whereas remaining zeros are considered as part of arising from the counting process and may be termed as high-risk zeros. ZI models typically use logistic model and a standard count model to describe non-counting and counting data respectively.

In ZI distribution, non-count model will provide probability of being a zero from a non-counting process whereas standard count model provides probability of being zero from a counting process. These probabilities can be used to estimate the proportions of low and high risk zeros. Thus, ZI models can be used not only to account for the variability due to excess zeros but also to estimate the mixing proportions of these latent zeros.

High-risk zeros are naturally considered to be part of the count process. Those subjects identified as such would be likely at higher risk of event of interest as compared to those at low/no risk. Thus, there is not only a need to estimate the proportion of such zeros in a cohort but also to correctly identify the observed zeros at higher risk. The predicted probabilities of zeros from non-count and standard count models using ZI models can be used to classify latent high-risk zeros. Various procedures can be adopted using these probabilities to classify latent class zeros. We propose and compare three simple methods to classify latent zeros (low and high) using simulations and a real study dataset.

2. Data

To demonstrate our proposed strategy, we have considered our motivating study data on nodal involvement in breast cancer patients. The details of the dataset are given in our previously published paper (Dwivedi et al., 2010). In this study, number of involved nodes was considered as outcome. This data set involved large proportion (38.8%) of patients with negative nodes. All patients were diagnosed with breast cancer so each patient was at risk of nodal involvement. Thus, zero hurdle models were more appropriate to account for the variability due to excess negative nodes. Since, due to disease process, some of the patients with negative node in the data set might have been observed as negative but may be at a high-risk of nodal involvement as compared to others with negative node. Further, it has been reported that some patients with the involved (positive) nodes may be recorded as negative due to misclassification by the pathologist or due to non-dissection of complete axilla. This clearly indicated that patients with negative node could be classified into a very low-risk and high-risk negative nodes. Thus, ZI models were used to account for the variability due to excess negative nodes and mixture of zeros. We demonstrated that the Zero Inflated Negative Binomial (ZINB) model fit and described the data well with number of involved nodes as outcome. In this dataset, it is a clear indication of involvement of high-risk negative nodes and so we needed to classify them as such so that either the miss-classification or future clinical management of the patients at high-risk can be handles efficiently. There is also a need to carefully follow the patients with high-risk negative nodes to prevent future risk of recurrence in these patients. We propose our strategy to classify patients into low and high-risk negative nodes using developed ZINB model.

3. Zero Inflated Model

Suppose Y is a random count variable with excess zeros. The zero inflated (ZI) count distribution of Y can be expressed as:

P(Y=0)= p+(1-p)*f(0) where y=0
P(Y=m)= (1-p)*f(m) where y=m; m=0,1,2,…

where f(.) is a distribution (Poisson/negative binomial) representing the count process and p is the probability of zero estimated through non-count process (logistic).

Note that the above ZI distribution of Y can be obtained by multiplying a binary random variable (Y1) with a count random variable (Y2) (Kelley and Anderson, 2008):

Y=Y1*Y2 Where Y1=0,1 and Y2= 0,1, 2……

$$
\begin{aligned}
P[Y=0] &= P[Y_1=0 \text{ and } Y_2=0]+P[Y_1=0 \text{ and } Y_2>=1]+ P[Y_1=1 \text{ and } Y_2=0] \\
&= pf(0)+p\{1-f(0)\} +(1-p)*f(0) \\
&= p+(1-p)*f(0) \qquad \text{when } y=0
\end{aligned}
$$

$$
\begin{aligned}
P[Y=m] &= P[Y1=1 , Y2=m] \\
&= (1-p)*f(m) \qquad \text{when } y>0
\end{aligned}
$$

Here, p denotes the probability of low-risk zero, f(0) is the probability of zero from count process, (1-p)f(0) denotes the probability of high-risk zero, and f(.) can be considered as Poisson/ negative binomial distribution. For nodal data set, ZINB was found to be appropriate thus, we use ZINB to demonstrate our proposed strategy.

If Y denotes the random variable denoting the number of involved nodes and if we estimate p from the logistic model and substitute f(.) as negative binomial distribution with then the distribution of Y follows ZINB model.

For the ith patient (i=1…n we write;

$$
p(y_i|x_i)= \begin{cases} p_i+(1-p_i)\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{1/\alpha} & , \ y_i=0 \\ (1-p_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{1/\alpha}\left(\dfrac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{y_i} & , \ y_i \geq 1 \end{cases} \qquad (1)
$$

Where pi indicates the probability of negative nodes in non-count process and fi(0) indicates probability of negative nodes and λ_1 is the mean of positive nodes in count process for the ith patient. The α(>0) represents the over-dispersion parameter due to unobserved heterogeneity. If γi's and βi's are the respective regression coefficients under logistic and negative binomial models corresponding to the considered covariates (xi's), and the number of considered covariates is k in each of the models, then using equation (1), regression models can be expressed as (note that it does not have to be same x's or the same number of x's)

$$\log\left(\frac{p_i}{1\text{-}p_i}\right) = \gamma_0+\gamma_1 x_1+\gamma_2 x_2+\text{---}+\gamma_k x_k \tag{2}$$

$$\log(\lambda_i) \quad = \beta_0+\beta_1 x_1+\beta_2 x_2+\text{----}+\beta_k x_k \tag{3}$$

Zero inflated models are mixture models that are often used with the datasets that contain large numbers of zeros. The conditions under which the zero inflated models are identifiable (or not) are already described in literature (Li, 2012, Cameron and Trivedi, 1998). If the observations are sufficient and observed information matrix is non-singular then there are no identifiability problems (Lambert, 1993).

4. Strategy for identifying high risk zeros

In ZI models, a proportion of zeros estimated through non-count model/logistic model is classified as low/no risk zeros (p) and remaining proportion of zeros is classified at high-risk zeros ((1-p)*f(0)). Total estimated zeros from the ZI models has the probabilityp+(1-P)*f(0). For each ith subject using ZI models, we can obtain using regression estimates and the cofactors xi's:

pi : Probability of being non count zero obtained from the non-count model.
fi(0) : Probability of being count zero obtained from the count model.

We explore three procedures (rank-based, likelihood-based, posterior probability–based) for classification discussed in the following sub-sections to classify zeros as low or high-risk zeros.

4.1 Rank based classification (RC)

The predicted probabilities of being zero from non-count process (p) or predicted probabilities of being zero from count process (f(0)) can be used to classify subjects with low and high risk zero. In this method, first we estimate the total probability of low risk zero (say p') in the study sample with observed zeros. After that, we sort the predicted probabilities of zero obtained either from logistic or count model and rank them according to their ascending or descending order among the observed zeros. We classify a proportion (p') of subjects as low risk with highest probability of low risk zeros. For nodal dataset, we used probability of being negative node estimated through logistic model in ZINB model to classify low and high-risk negative nodes.

4.2 Likelihood based classification (LC)

The likelihood of being low-risk zero (pi) and high-risk zero [(1- pi)* fi(0)] obtained from zero inflated model can be compared for each subject. If a subject observed with zero and if pi>[(1- pi)* fi(0)] then the subject is classified as at low-risk zero otherwise at high-risk zero. We compared the probability of being negative node estimated from logistic model with the probability of being negative node estimated from negative binomial model given that it is not from the logistic model. Given the patient with negative node, if the probability of being high-risk negative node was found to be higher than the probability of being low-risk negative node then patient is classified as high-risk negative node.

4.3 Posterior Probability Based Classification (PC)

We know the likelihoods of low-risk zero from non-count process (pi) and high-risk zero (1-pi)*fi from the count process for each ith subject thus we can obtain posterior probabilities of being from the two latent classes.

We can use observed proportions of zero and non-zero as prior probabilities or weights for count and non-count processes. Classifying zeros using posterior probabilities are equivalent with classifying zeros with weighted probability of each process. Suppose, T and (1-T) are the observed proportions of zeros and non-zeros respectively in the dataset then:

Weighted probability of a subject with low risk zero = pi*T
Weighted probability of a subject with high risk zero = [(1- pi)* fi(0)] *T

Given the zero, if weighted probability of high risk zero is larger than the weighted probability of low risk zero then the subject would classify as at high risk zero otherwise at low risk zero. In the nodal dataset, for a node-negative patient, weighted probabilities of being at low-risk negative node and being a high-risk negative node were compared. We then assign subjects to one or the other based on the higher of the two probabilities.

To examine the risk levels of subjects with high risk negative nodes as compared with low risk negative nodes after adjusting positive nodes, a multinomial logistic regression was used. Suppose p1 is the probability of high risk zero, p2 is the probability of low risk zero and p3 is the probability of presence of positive nodes then multinomial logistic regression can be expressed as:

$$\log\left(\frac{p_1}{p_2}\right) = b_0 + b_1 x_1 + b_2 x_2 + \text{---} + b_k x_k$$

where b0 is the intercept and b1 to bk are the regression coefficients respected to k independent cofactors (x).

5. Monte Carlo simulations

To validate the classification approaches used for negative nodes classification, we carried out simulation studies to find answer for the following two questions:

1. Does ZI model correctly estimate proportion of low and high-risk zeros in a dataset
2. Which of the three approaches (RC, LC or WC) can be used to classify subjects into high and low-risk zeros

We simulated 1000 observations from ZINB distribution on a-priori assumption of low-risk zeros (p=27%) and rate of count events (4.01) as found in our real dataset. To simulate data from ZINB distribution, we first generated two random variables with 1000 observations. One variable from a Bernoulli (Y1) distribution and another from a negative binomial (Y2) distribution using a cofactor (X) following a standard normal distribution with regression coefficient B1 and B2 were generated. For the sake of simplicity, same variable (X) was considered for simulating two distributions. Multiplying two random variables Y1 and Y2 gives ZINB distribution. Further, different associations of a covariate (X) with Y1 and Y2 may affect the accuracy of proposed classification procedures. Thus, we created four simulated datasets corresponding to each of the following four regression situations:

1. X is highly associated (B1 and B2=0.9) with Y1 and with Y2 (First Regression : FR)
2. X is moderately associated (B1 and B2=0.5) with Y1 and with Y2 (Second Regression : SR)
3. X is moderately highly associated (B1=0.9) with Y1 and mildly associated (B2=0.1) with Y2(Third Regression : TR)
4. X is mildly associated (B1=0.1) with Y1 and highly associated (B2=0.9) with Y2(Forth Regression : FTR)

The observed zeros from Bernoulli random variable (Y1) or zeros from both Bernoulli random variable (Y1) and negative binomial random variable (Y2) are considered as true low-risk zeros. However, the observed zeros from negative binomial random variable (Y2) when Bernoulli random variable (Y1) is not zero are considered as true high-risk zeros. Thus, the subjects with true low risk and high risk zero are known in each simulated dataset. In each simulated dataset, a ZINB model can be fitted to estimate the proportion of total zero including low and high risk zeros and to compare with the true proportions of low and high risk zeros.

To address the question i.e., ZI model correctly estimates proportion of low and high-risk zeros in each dataset, we fit ZINB regression for each of the four regression conditions on a set of 1000 observation. We estimate low and high-risk zeros using ZI model for each regression condition. We then replicate the process on 100 simulated datasets under each regression condition and estimate the average proportion of low and high-risk zeros. Thereafter, we compare estimated average low and high-risk zeros with average true low and high-risk zeros.

To address the question i.e., RC, LC or PC approaches can be used to classify subjects into high and low-risk zeros, we fit ZINB regression for each of the four regression equations as above, and used the three procedures (RC, LC and PC) to classify low and high-risk zeros under each regression condition. We then replicate the process of classification on 100 simulated datasets under each regression condition. The estimated low-risk and high-risk zeros with classified low and high-risk zeros using each of the three procedures under each of the regression condition were compared. We also estimate sensitivity, specificity and diagnostic accuracy of each of three procedures in relation to the true zeros. The sensitivity (Se) is calculated as the proportion of high risk zero classified by an approach given the true high risk zero. The specificity(Sp) is defined as the proportion of low risk zero classified by an approach given the true low risk zero. The positive predictive value (PPV) is computed as the proportion of true high risk zero among the high risk zero classified by an approach. The negative predictive value (NPV) is computed as the proportion of true low risk zero among the low risk zero classified by an approach. Accuracy of the procedure is computed as the proportion of correctly identified low and high risk zeros.

6. Results

Of the 1152 patients, 38.8% patients were observed with negative nodes (zeros). ZINB model predicted that 38.6% negative nodes and 27.4% as low risk negative nodes in the data set. Thus, ZINB model predicted that 70.6% of all negative nodes were at "low risk" zeros, and the remaining 29.4% of the negatives nodes were at "high risk" for nodal involvement. Figure 1 compares the classification of negatives nodes using LC and PC with RC (estimated by ZINB model). LC classified 44 (9.8%) of the observed negative nodes at high-risk negative nodes and remaining at low-risk negative nodes. However, 22.6% of the observed negative nodes are classified at high-risk negative nodes using PC approach given the observed negative nodes. It seems that LC method underestimates the high-risk negative nodes for this datasets. PC method provides better classification than LC method. Proportion of concordance pairs between PC with other two methods was found to be more than 80%.

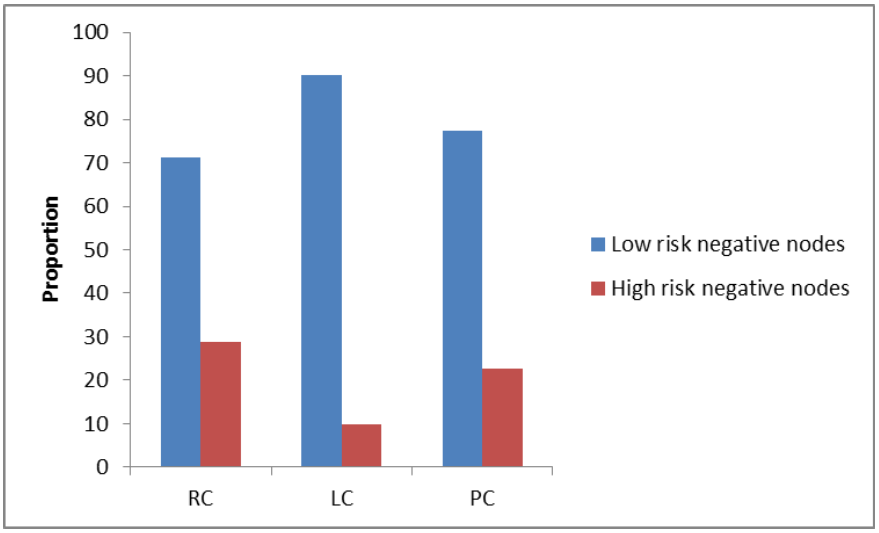

Figure 1. Comparison of estimated and classified negative nodes using different classification approaches

To compare the performances of each of the methods in classifying high-risk zeros, simulations studies were carried out. Figure 2 reveals the results of simulation studies. In any of the four regression situations, the ZINB model accurately estimates low and high risk zeros. Slightly high variation is observed in estimating low and high-risk zeros in case of fourth regression condition (i.e., less associated with non-count and highly associated with count process). Figure 3 reveals the classification of estimated high-risk zeros using the three methods among the observed zeros for four regression conditions. As obvious, RC method classifies the exact estimated high-risk zeros. LC method under classifies the estimated high-risk zeros in any of the four regression situations except first regression situation when variable is highly associated with both processes. However, PC method classifies high-risk zeros appropriately in any condition. The absolute bias in classifying high-risk zeros was less than 5% except third regression condition. In other words, if cofactor is highly associated with non-count process and less associated with count process then PC method ,on an average, under-classifies high-risk zeros. In any situation, PC method works better than LC method in classifying estimated high-risk zeros.

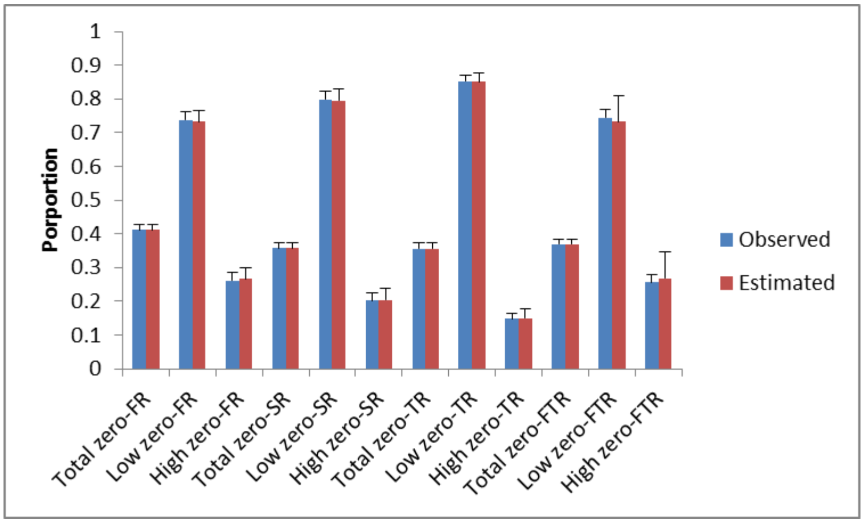

Figure 2. Comparison of observed and estimated zeros using zero inflated negative binomial model under four regression situations

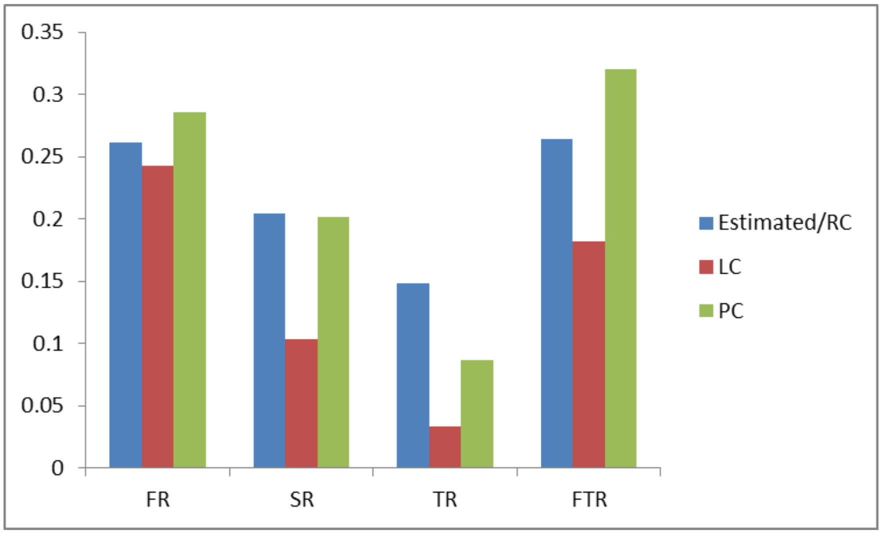

Figure 3. Comparison of estimated and classified zeros using different classification approaches under four regression situations

Overall diagnostic accuracy of correctly identifying the high-risk zeros was lower for RC method in any regression situation (Table 1). Sensitivity of classifying true high-risk zeros was lower for LC method. PC method was more sensitive method than LC method. All of the approaches had specificity more than 75%. LC method was found to be highly specific method than PC and RC methods. Each method has almost similar diagnostic ability in first regression scenario. Thus, any approach can be used in first regression situation. In second regression situation, performances of RC and PC were similar. LC method has very lower sensitivity for classifying high-risk zero in this case. Thus, PC or RC method should be preferred in second regression situation. For third regression situation, PC method has lower sensitivity than RC method but higher sensitivity than LC method. Also, other indicators of PC method were better than RC method and similar to LC method. Keeping in view of additional advantages of PC over RC, in this situation, again PC should be preferred. The diagnostic ability of RC method was found to be highly variable than the PC method especially for fourth regression situation. PC method should be preferred in case of fourth regression situation. In summary, it indicates that PC method should be preferred in comparison to RC and LC methods in any regression situation. The proportion of concordant pairs among any of the three procedures was more than 85%. Proportion of concordant pairs was higher between PC and RC methods for any regression situation. Thus, in general, PC method should be preferred over other two methods.

Table 1. Diagnostic performance of the different classification approaches for classifiying high risk zeros

			FR		**SR**		**TR**		**FTR**	
		N	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
RC	Se	100	0.72	0.06	0.52	0.08	0.40	0.08	0.48	0.20
	Sp	100	0.89	0.02	0.87	0.03	0.89	0.02	0.81	0.09
	PPV	100	0.71	0.05	0.51	0.06	0.40	0.07	0.46	0.18
	NPV	100	0.89	0.02	0.82	0.03	0.82	0.02	0.76	0.11
	Accuracy	100	0.85	0.02	0.80	0.02	0.82	0.02	0.72	0.11
LC	Se	100	0.68	0.08	0.30	0.12	0.13	0.09	0.39	0.14
	Sp	100	0.91	0.03	0.95	0.03	0.98	0.02	0.89	0.10
	PPV	100	0.73	0.05	0.61	0.10	0.57	0.19	0.58	0.07
	NPV	100	0.89	0.03	0.84	0.03	0.87	0.02	0.81	0.07
	Accuracy	100	0.85	0.02	0.82	0.02	0.86	0.02	0.76	0.05
PC	Se	100	0.75	0.08	0.50	0.12	0.26	0.13	0.62	0.12
	Sp	100	0.88	0.03	0.87	0.05	0.94	0.04	0.78	0.10
	PPV	100	0.69	0.05	0.51	0.08	0.46	0.13	0.51	0.05
	NPV	100	0.91	0.02	0.88	0.03	0.88	0.02	0.85	0.07
	Accuracy	100	0.85	0.02	0.80	0.03	0.84	0.02	0.74	0.05

Table 2. Factors differentiating low and high-risk negative nodes after adjusting positive nodes using multinomial logistic regression

Variables	**Low risk vs. High risk negative nodes**	
	OR (95%CI)	**p-value**
Age (year)		
>35		
<=35	0.74(0.33, 1.65)	0.464
Symptom duration (months)		
<=2		
3-4	3.56(1.78, 7.12)	0.000
5-8	0.38(0.16, 0.91)	0.029
>=9	1.85 (0.84, 4.07)	0.126
Parity		
Nulliparous		
P1/P2	0.08(0.02, 0.28)	0.000
Multiparous	0.01(0.00, 0.04)	0.000
Menopausal		
Post Menopausal		
Pre Menopausal	4.47(2.44, 8.18)	0.000
Primary side		
Left		
Right	1.74(1.01, 2.99)	0.046
Primary site		
Medial (UIQ+LIQ)		
Lateral (LOQ+UOQ)	3.12(1.38, 7.05)	0.006
Central/Multiple/ Other	14.69(5.70, 37.86)	0.000
Skin changes		
No		
Yes	7.69(3.82, 15.50)	0.000
Tumor type		
Other /ILC		
IDC	1.39(0.48, 4.01)	0.545
Tumor size (centimeter)		
<=2		
2-5	3.75(1.81, 7.76)	0.000
>5	1.12(0.44, 2.87)	0.812
Neoadjuvant chemo		
No		
Yes	0.85(0.38, 1.90)	0.693
Total dissected nodes	0.75(0.69, 0.80)	0.000

Since PC was found to be better approach in classifying high risk zeros, we use PC method to classify high-risk negative nodes in nodal dataset to assess the risk level between classified low risk negative nodes, high risk negative nodes and positive nodes. Table 2 reveals the results of multinomial logistic regression to assess the risk level between low-risk negative nodes and high-risk negative nodes after adjusting positive nodes. Patients reported 3-4 months symptom duration were more likely to have high-risk negative nodes. If a patient observed with negative node and more than or equal to single parous then that patient was more likely to be a low-risk negative node. Pre-menopause status leads to high-risk negative nodes. Patients observed with central or multiple sites tumor more were more likely to be at high-risk negative nodes. Skin positive patients had more chance of high-risk negative nodes. In summary, this analysis reveals that patients observed with negative nodes with skin changes, multiple/central tumor site, 3-4 months of symptom duration, pre-menopausal status and nulliparous status significantly increase the likelihood of at high-risk negative nodes than low risk negative nodes.

7. Discussion

Developing clinical decision rules to classify new patients into various clinically important categories are very common in clinical research using classification and regression trees or regression approaches (Lewis, 2000). These rules help in classifying and identifying new patients who are at risk of certain event of interest. However, it also becomes important to classify objectively those patients who are already diagnosed with disease free state but more at risk to develop the disease in near future. As a standard medical care, clinicians usually use the set protocol to follow up and re-examine the disease of all negatively diagnosed patients. In other words, clinicians usually treat all negatively diagnosed patient at equal risk level and sometime subjectively to decide their follow up times and re-examinations. However, when outcome is measured through degree of disease (i.e., count or ordinal) involving predominately no disease/event of interest, which is more common in practice, then there is likelihood of some patients with relatively high-risk of disease who are still in disease free state. Thus, it becomes important to classify negatively diagnosed patients who are at high-risk of disease/ event as objectively as possible according to presence/absence of risk factors including at what degree.

In nodal data sets, almost 30% negative nodes estimated to be at high-risk for nodal positivity. This indicates that a significant proportion of patients with negative nodes need to be re-examined or followed up more closely to minimize the risk of event in near future. Posterior based approach identifies 23% of the negative nodes at high-risk. This indicates that a large proportion (77%) of estimated high-risk negative nodes could be identified through posterior based approach. Likelihood based approach could identify only 33% of the estimated patients with high-risk negative nodes. This indicates that PC approach works better than LC approach in identifying at high-risk negative nodes. PC approach differs only with LC approach in terms of prior probability. Here, we assumed the prior probabilities of latent classes (i.e., non-count and count processes) as total number of observed negative nodes and positive nodes respectively. Good estimates of prior probabilities may improve further classification of high-risk negative nodes. Obviously, rank based method identifies all estimated high-risk negative nodes because it is conditioned on estimated high-risk negative nodes. This approach can only be used to classify a group of patients not to classify an individual patient. There is high concordance observed in identifying high-risk negative nodes between RC and PC methods indicating both can be used to classify.

Estimation and classification of zeros should depend on number of cofactors used to develop the regression model as well as their level of association in predicting outcome. Thus, we considered four types of regression situations to evaluate the performance of estimating and classifying low and high-risk zeros using ZINB in case of excess zeros. In any of the situations, analysis reveals that ZINB model accurately estimates the low and high-risk zeros. ZINB provided slightly high variation in the estimation of high risk zeros when regression coefficient is highly associated with count process and mildly associated with non-count process. This may happen because ZI models are used to account for excess variability due to zeros in non-count process (Ridout et al., 1998). When variable is mildly associated with non-count process then this may not be able to explain complete variability due to excess zeros subsequently may underestimate the low-risk zeros.

LC method under-classifies the high-risk zeros in any condition. This seems reasonable because, it does not account for the prior probabilities of latent classes. PC method has been used to classify latent classes in latent class modeling (Proust and Jacqmin-Gadda, 2005). PC method under-classifies high- risk zeros when cofactor is less associated with count process and highly associated with non-count process. Classification approaches highly depend on the cofactors abilities to predict the latent classes. Thus, if cofactor ability is very weak to predict count process zeros then it will under classify count zeros. In our nodal data sets, all the significant cofactors for non-count process are less associated (less than regression coefficient .2) and highly associated with count process (more than .9). This is an example of third regression situation in simulation studies. This indicates that PC may under classify high-risk zeros in such case. We also found that PC method slightly under classifies negative nodes for nodal dataset.

We showed in simulation studies that PC method performs well for classifying high-risk zeros as estimated through ZINB model and also it correctly identifies the high-risk zeros. The accuracy for correctly identifying high and low-risk zeros was found to be more than 80% in any regression situation and with any approach. The sensitivity of classifying high-risk zeros was found to be high in all the conditions except third regression situation. For this regression situation, there is a need to examine some other approaches that can provide better classification and identification of high-risk zeros than PC, and RC approaches.

Although RC method works well for nodal dataset, in view of classifying individual using PC method, PC method was used to asses cofactors associated with high risk zeros. If a patient is diagnosed with nodal negativity then there is a need to focus on cofactors such as presence of skin changes, multiple/central tumor site, 3-4 months of symptom duration, pre-menopausal status. Patients with these cofactors need to follow up more closely and re-examine for confirmation of no nodal involvement. Patients with absence of such factors may need to have extended follow up period and may be avoided for further more diagnostic examinations to reduce cost burden on the patients.

Generally, the appropriate length of follow up period depends on the duration of disease, disease process and presence of risk factors. The developed procedure in this study can be used for many purposes such as selecting screening population for certain disease, updating the risk evaluation of negatively diagnosed patients, validating the confirmatory diagnosis of the disease and setting up follow up period and re-examination times for negatively diagnosed patients. This procedure utilizes the existing information to set up rules for future management of the negatively diagnosed patients. To use this procedure, there is need to first review the history of the patients whose clinical outcome/ disease severity is measured through count or ordinal data involving large proportion of no outcome of interest. Appropriate ZI models for count or ordinal data need to be developed and validated using important cofactors. Newly negatively diagnosed patients can be classified at high-risk and low-risk of disease for their future management using PC method. Also, this procedure can be used to select screening population who are at risk of disease to carry out studies on risk patients.

There are some limitations of this study. We used nodal dataset to demonstrate our proposed strategies to estimate and classify low and high negative nodes patients. Some of the important cofactors could not be included for the development of ZINB model. Inclusion of such cofactors may improve classification accuracy of the procedures. We assessed classification accuracy of the proposed methods using single continuous cofactor in simulation studies. Inclusion of a list of cofactors may influence the diagnostic performances of the various approaches. We used RC method using probability of low risk zeros. The RC approach can be used using probability of high-risk zeros or probability of count zeros or probability of total zeros. This may change the diagnostic performance of the RC approach. Simulation studies were carried out using a-priori assumption of low-risk zeros (p=27%) and rate of count events (4.01) as found in our real data set. The performances of various diagnostic procedures in other scenarios can also be explored.

8. Conclusion

ZI models can be used accurately to estimate low and high risk zeros. PC method can be used to classify low and high-risk zeros appropriately in most of the regression situations. A significant proportion of nodal negative patients could be followed up more closely to avoid risk of events and re-examined for confirmation of nodal positivity. The developed procedure can be used to update, validate and monitor the newly diagnosed nodal negative cases. The approach developed here is applicable to other diagnostic scenarios where the disease or outcome can be characterized by count/ordinal data involving high proportion of zeros.

9. Acknowledgements

This work was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 8 UL1 TR000077-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Cameron, A.C., and Trivedi P.K. (1998). *Regression Analysis of Count Data*, Cambridge, MA: Cambridge University Press.

Dwivedi, A.K., Dwivedi, S.N., Deo, S.V.S., Shukla, R. and Kopras, E. (2010). Statistical models for predicting number of involved nodes in breast cancer patients. *Health* 2(7): 641-651.

Kelley, M.E., and Anderson, S. J. (2008). Zero inflation in ordinal data: Incorporating susceptibility to response through the use of a mixture model. *Statistics in Medicine* 27(18): 3674–3688.

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

Lewis, R.J. (2000). An introduction to classification and regression tree (CART) analysis. *Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco*, California.

Li, C.S. (2012). Identifiability of zero inflated Poisson models. *Brazilian Journal of probability and Statistics* 26(3):306-312.

Mohri, M. and Brian R. (2005). Structural zeros versus sampling zeros. *Technical Report CSEE*-05-003,

Proust, C., Jacqmin-Gadda, H. (2005). Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer Methods and Program in Biomedicine* 78:165-173.

Ridout, M., Demetrio, C.G.B., Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*: 1-13.

Rose, C.E., Martin, S.W., Wannemuehler, K.A. and Plikaytis, B.D. (2006). On the use of zero- inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics* 16(4), 463–481.

Slymen, D.J., Ayala, G.X., Arredondo, E.M. and Elder, J.P. (2006) A demonstration of modeling count data with an application to physical activity. *Epidemiologic Perspectives & Innovations*, 3, 3.

Received April 3, 2013; accepted September 30, 2013.

Assistant Professor Alok Kumar Dwivedi,
Texas Tech University Health Sciences Center
El Paso, Texas, USA
Tel: 915-215-4177
Fax: 915-545-5716
E-mail: alok.dwivedi@ttuhsc.edu

Professor Sada Nand Dwivedi,
All India Institute of Medical Sciences
New Delhi. India
E-mail: dwivedi7@hotmail.com

Professor S.V. S. Deo,
All India Institute of Medical Sciences
New Delhi. India
E-mail: svsdeo@yahoo.co.in

Professor MB Rao,
Center of Biostatistical Services, College of Medicine
University of Cincinnati, Cincinnati-45267, Ohio, USA
E-mail: raomb@ucmail.uc.edu

Professor Rakesh Shukla,
Center of Biostatistical Services, College of Medicine
University of Cincinnati, Cincinnati-45267, Ohio, USA
E-mail: shuklar@ucmail.uc.edu