Published in final edited form as:

Stat Biopharm Res. ; 4(3): 252–263. doi:10.1080/19466315.2011.636279.

The Zero-truncated Poisson with Right Censoring: an Application to Translational Breast Cancer Research

Hung-Wen Yeh, PhD^{1,*}, Byron Gajewski, PhD^{1,2}, Purna Mukhopadhyay, PhD³, and Fariba Behbod, PharmD, PhD⁴

¹Department of Biostatistics, The University of Kansas Medical Center, Kansas City, Kansas 66160

²School of Nursing, The University of Kansas Medical Center, Kansas City, KS 66160

³Arbor Research Collaborative for Health, Ann Arbor, MI 48104

⁴Department of Pathology and Laboratory Medicine, The University of Kansas Medical Center, Kansas City, KS 66160

Abstract

We propose to analyze positive count data with right censoring from Behbod et al. (2009) using the censored zero-truncated Poisson model (CZTP). The comparison in truncated means across subgroups in each cell line is carried out through a log-linear model that links the un-truncated Poisson parameter and regression covariates. We also perform simulation to evaluate the performance of the CZTP model in finite and large sample sizes. In general, the CZTP model provides accurate and precise estimates. However, for data with small means and small sample sizes, it may be more proper to make inference based on the mean counts rather than on the regression coefficients. For small sample sizes and moderate means, the likelihood ratio test is more reliable than the Wald test. We also demonstrate how power analysis can be used to justify and/or guide the choice of censoring thresholds in study design. A SAS macro is provided in Appendix for readers' reference.

Keywords

positive counts; zero-truncated Poisson; right censored; small sample size; power; ductal carcinoma in situ

1. Introduction

The evolution of breast cancer involves the development of non-invasive lesions followed by invasion and metastasis. Ductal carcinoma *in situ* (DCIS) is a common form of human breast non-invasive lesions. Research indicates heterogeneous potential of progression among DCIS patients (Boughey et al., 2007) and the heterogeneity may arise early at the non-invasive stages in DCIS (Allred et al., 2008). Hence, the knowledge about the heterogeneity in progression potential is crucial as it enables early detection and identification of patients who are more likely to progress to more severe stages. One approach to address the issue is using animal models (Shackleton et al., 2006; Behbod et al., 2009).

The problem that we consider in this paper comes from the study of Behbod et al. (2009) who conducted a human-to-mouse intraductal transplantation to mimic natural progression of

---

*Corresponding author hyeh@kumc.edu Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160 Tel: 1-913-588-0531 Fax: 1-913-588-0252 .

human DCIS and to study heterogeneity of the disease in mice. One of the outcome measures was the number of mammary ducts with DCIS-like lesion. However, the sample sizes were small and the counts included right censored observations. Our interest is to evaluate and compare the number of affected ducts which reflects the progression of the disease.

This article is composed as follows. We describe the problem in Section 2. In Section 3, we revise the established ZTP model (Xie and Aickin, 1997) for count data with simultaneous left-truncation and right-censoring. The Poisson parameter is fitted by the log-linear model and estimated by the maximum likelihood method. In Section 4, we analyze the data and conduct a simulation study to evaluate the accuracy and precision of maximum likelihood estimates (MLE). The simulation of standard ZTP without censoring has been discussed ([R] simulated zero-truncated Poisson distribution). Here, we provide a SAS macro (see Appendix) to simulate and analyze positive counts with right censoring. Finally, some concluding remarks and discussion are given in Section 5.

2. Problem description

Behbod et al. (2009) adopted two established human DCIS cell lines, “MCF10DCIS.COM” (abbreviated as DCIS.COM thereafter) and “SUM-225”, sorted a fixed number of cells from each subpopulation according to the expression of certain mammary progenitor surface markers (CD24/CD44, CD49f/CD24, and CD49f/MUC-1). The surface marker expression level was arbitrarily designated as low (lo) if the log₁₀ of the median fluorescence intensity (FI) was between 0 and 2, medium (med) if the FI was between about 2.1 and 3.6, and high (hi) if the FI was higher than 3.7. The cells from each subpopulation were sorted and evaluated for in vivo growth potential by intraductal injection. For example, a group of mice that were injected with human cells expressing hi level of CD44 and low level of CD24 was noted as CD44^{hi}/CD24^{lo} (see Table 1), and other groups of mice were named in the similar manner. The choice of the experimental groups was based on previous literature (citation 10 and 11 in Behbod et al., 2009). An additional group of mice (the “random sort” in Table 1) was treated with human cells sorted arbitrarily (contrast to using the surface markers) as a control group. The sample size was determined by the number of human cells available following flow cytometry sorting. A fixed number of 5,000 DCIS cells were used for each group of mice. For the “random sort” group, the 5,000 cells were enough for 28 mouse fat pads (Table 1, Column # injections). When the 5,000 cells were sorted according to CD49f/MUC-1, for instance, the cells expressing hi level CD49f and medium level MUC-1 (CD49f^{hi}/MUC-1^{med}) or medium CD49f and low MUC-1 (CD49f^{med}/MUC-1^{lo}) were enough for 23 injections, but those expressing hi level CD49f and low level MUC-1 (CD49f^{hi}/MUC-1^{lo}) or those expressing medium level in CD49f and MUC-1 (CD49f^{med}/MUC-1^{med}) could only afford 6 fat pads. At six weeks after intraductal injection, the researchers examined whether positive growth appeared (i.e. DCIS-like lesions developed) in animal fat pads (fps) and if so, they counted the numbers of ducts with such lesions. However, counts greater than seven were recorded as “7+” thus created right censored observations.

Although the counts of ducts with DCIS-like lesions do include zeros, the zero counts were due to two possibilities: lack of progenitor cells (Shackleton et al., 2006, where MRUs are progenitors) or some technical issues that caused progenitor cell death in the process of transplantation. The researchers had carefully examined cell survival prior to transplantation thus the second reason was not likely. When progenitor cells exist, the transplant would grow and lesions would develop within 6 weeks. If a DCIS-like lesion was not observed by 6 weeks after transplantation, it was unlikely that any lesion would appear at a later time point. In the original paper of Behbod et al., the authors compared across subgroups the proportions of positive growth (non-zero counts), which indicates the existence of tumor progenitor cells. Here, we would like to estimate and compare the positive numbers of ducts with DCIS-like

lesions, which indicates progression potential of human DCIS cells. Combining the zero and positive counts can veil the real progression potential. For instance, a type of cells may contain fewer progenitor cells and shows lower proportion of positive growth but the progenitor cells, if exist, may be more aggressive and cause lesion in more ducts. In this case, combining the zero and positive counts may lead to a lower “apparent” means. In this content, the zero counts may be considered as the results of “failed” experiments and are excluded. Therefore, only the positive counts are analyzed here so that we can distinguish the scientific question of progression potential from that of the existence of cancer progenitors. Note that some subgroups from the original paper are excluded to avoid computational problems because they either have identical observations or do not have enough non-censored observations. We perform the analysis using the zero-truncated Poisson (ZTP) with right censoring taken into account.

3. The censored zero-truncated (CZTP) model

Let Y denote a positive count that follows the ZTP distribution from the parent Poisson distribution with parameter λ . Then the probability mass function is given as

$$f(y) = \Pr(Y=y|Y>0) = \frac{\lambda^y}{(e^\lambda - 1)y!}, \quad y=1, 2, 3, \dots \quad (1)$$

With the aid of factorial moments, the conditional mean and the conditional variance can be expressed in terms of the unconditional mean λ and an adjustment factor δ (see Gurmu, 1991):

$$E(Y|Y>0) = \mu = \lambda / (1 - e^{-\lambda}) = \lambda + \delta \quad (2)$$

$$\text{Var}(Y|Y>0) = \sigma^2 = \mu(1 - \delta) = (\lambda + \delta)(1 - \delta) \quad (3)$$

where $\delta = \lambda/(e^\lambda - 1)$, and ranges between 0 and 1. Thus in Equations (2) and (3), $\mu > \lambda$ and $0 < \sigma^2 < \mu$, i.e. the conditional mean is greater than the unconditional mean of the parent Poisson distribution, and the conditional variance is less than the unconditional variance. Moreover, as $\lambda \rightarrow \infty$, $\delta \rightarrow 0$ and the ZTP distribution approaches the parent Poisson distribution.

In the motivating study, censored observations exist. Suppose among the $N = n + m$ observations, n of them are exactly counted and the rest m observations are right censored ($Y_i = h_i$, $i = n + 1, n + 2, \dots, N$). Then the likelihood function can be written as

$$L(\lambda) = \prod_{i=1}^N \left[\frac{\lambda^{y_i}}{y_i! (e^\lambda - 1)} \right]^{1-c_i} [\Pr(Y_i \geq h_i)]^{c_i}$$

where c_i is a dummy variable indicating whether the i^{th} observation Y_i is censored at the threshold h_i :

$$c_i = \begin{cases} 1 & \text{if } Y_i \geq h_i \\ 0 & \text{if } Y_i < h_i \end{cases}$$

The log-likelihood and its first and second derivatives are then

$$l(\lambda) = \log L(\lambda) = \sum_{i=1}^n \left[y_i \log \lambda - \log(y_i!) - \log(e^\lambda - 1) \right] + m \log [\Pr(Y \geq h)]$$

$$\frac{dl(\lambda)}{d\lambda} = \frac{1}{\lambda} \left[n(\bar{y}_n - \mu) + m \frac{\sum_{j=1}^{h-1} f(j)(\mu - j)}{\Pr(Y \geq h)} \right]$$

$$\frac{d^2 l(\lambda)}{d\lambda^2} = -\frac{1}{\lambda^2} \left\{ n(\bar{y}_n - \mu + \sigma^2) + m \left[\left(\frac{\sum_{j=1}^{h-1} f(j)(\mu - j)}{\Pr(Y \geq h)} \right)^2 - \frac{\sum_{j=1}^{h-1} f(j) [(\mu - j)^2 + \sigma^2 - (\mu - j)]}{\Pr(Y \geq h)} \right] \right\}$$

where $\bar{y}_n$ is the mean of the n observations that have an exact measurement.

To find maximum likelihood estimate by solving λ , we set $dl(\lambda)/d\lambda$ equal to 0. Since this is a non-linear function, one can use numeric methods such as the Newton-Raphson algorithm to find $\hat{\lambda}$, and then obtain $\hat{\mu}$ and $\hat{\sigma}^2$ using Equations (2) and (3) by the invariance property. Note that when there is no censoring, $m = 0$ and $n = N$, hence $\hat{\mu} = \bar{y}_n = \bar{y}_N$. Variance of $\hat{\mu}$, can be estimated via the variance of $\hat{\lambda}$, which is obtained by the negative reciprocal of $d^2 l(\lambda)/d\lambda^2$ evaluated at $\lambda = \hat{\lambda}$.

Consider p covariates and let $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \dots, x_{p,i})'$ be the covariate vector for the i^{th} observation ($i = 1, 2, \dots, N$), and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ be the associated parameter vector. We then link the parameter λ and regression coefficients by the log-linear regression model

$$\log \lambda_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} = \mathbf{x}_i' \boldsymbol{\beta} \quad \text{or} \quad \lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}) \quad (4)$$

The DCIS.COM experiment contains 9 groups. Using the random sort as the reference group, we need 8 dummy variables, with $x_{j,i} = 1$ if observation i is from group j and 0 otherwise, $j = 1, 2, \dots, 8$ (see Table 1 for group labeling); on the other hand, we only consider two groups of SUM-225, and let $x_{1,i} = 1$ if observation i is from the treatment of CD49^{med}/CD24^{hi} and 0 otherwise.

Then the log-likelihood and the first two derivatives are expressed as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^N \left\{ (1 - c_i) [y_i \log \lambda_i - \log(y_i!) - \log(\exp(\lambda_i) - 1)] + c_i \log \Pr(y_i \geq h_i | y_i > 0) \right\}$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \left\{ (1 - c_i) (y_i - \mu_i) - c_i \left[\frac{\sum_{\ell=1}^{h_i-1} \Pr(Y_i = \ell | Y_i > 0) (\ell - \mu_i)}{\Pr(Y_i \geq h_i | Y_i > 0)} \right] \right\} \mathbf{x}_i$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = - \sum_{i=1}^N \left\{ (1 - c_i) \sigma_i^2 + c_i \left[\left(\frac{\sum_{\ell=1}^{h_i-1} (\ell - \mu_i) \Pr(Y_i = \ell | Y_i > 0)}{\Pr(Y_i \geq h_i | Y_i > 0)} \right)^2 + \frac{\sum_{\ell=1}^{h_i-1} [(\ell - \mu_i)^2 - \sigma_i^2] \Pr(Y_i = \ell | Y_i > 0)}{\Pr(Y_i \geq h_i | Y_i > 0)} \right] \right\} \mathbf{x}_i \mathbf{x}_i'$$

where σ_i^2 is the conditional variance for the i^{th} observation (Equation (3)). Finally, the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$, $\mathbf{I}(\hat{\boldsymbol{\beta}})$, is simply the negative inverse of the second derivative evaluated at $\hat{\boldsymbol{\beta}}$.

The global effects can be examined by the likelihood ratio test (LRT) on the null hypothesis about the truncated means $H_0: \mu_0 = \mu_1 = \dots = \mu_p$ using the log-likelihood functions of the

full and the null models. The test statistic $G^2 = -2 \log \Lambda = -2 [l(\hat{\beta}) - l(\beta_0)]$, where β_0 is a null vector, asymptotically follows a χ^2 distribution with p degrees of freedom. Note that β 's are used to incorporate multiple groups or covariates and to estimate the un-truncated as well as the truncated means, but the likelihood function is based on the ZTP distribution hence the LR test still assesses the homogeneity in *truncated* mean. If $H_0: \mu_0 = \mu_1 = \dots = \mu_p$ is rejected, one may also perform post hoc comparisons on $H_0: \mu_i = \mu_j$ using $z = (\hat{\mu}_i - \hat{\mu}_j) / SE(\hat{\mu}_i - \hat{\mu}_j)$ where $\hat{\mu}_k = \hat{\lambda}_k / (1 - e^{-\hat{\lambda}_k})$, $k = 1, 2, \dots, p$, and $\hat{\lambda}_0 = \exp(\hat{\beta}_0)$, $\hat{\lambda}_\ell = \exp(\hat{\beta}_0 + \hat{\beta}_\ell)$, $\ell \neq 0$. Testing on $H_0: \beta_j = 0$ is not proper because the regression coefficients are linked to the *un-truncated* means whereas our inference focuses on the *truncated* means.

4. Results and Simulation

Overall, the data for the DCIS.COM and the SUM-225 cell lines include $2/43 \cong 5\%$ and $4/10 = 40\%$ censored observations, respectively (Table 1). For each cell line, we fit the data using the CZTP model. To apply this method, we assume that (1) development of DCIS-like lesions was independent of one another, (2) lesion-developing rate was not exponential, (3) each duct had the same probability of being affected, (4) if injected cells includes no progenitor cells, no lesion would appear, and (5) when the injected cell culture contains progenitor cells, lesions would develop in one or more mammary ducts within 6 weeks.

Table 1 shows the observed positive counts, and mean counts (μ in Equation (2)) and standard error (SE). The MLE of regression coefficients in the log-linear model for the un-truncated Poisson mean (Equation (4)) and the corresponding SE are given in Table 2. The LR test suggests the difference in truncated means across DCIS.COM subgroups is not significant at 0.05 level (LRT statistic $G^2 = 14.2$ and $p = 0.077$). For SUM-225, the between-group difference is not significant, either (LRT statistic $G^2(1) = 0.2$ and $p = 0.65$). The model goodness-of-fit is evaluated by the Pearson's Chi-square test. We combine categories of the counts 5 or greater in DCIS.COM such that the expected frequency in each category is at least 5, and find no evidence the model does not fit the data ($\chi^2(4) = 1.07$ and $p = 0.90$). The SUM-225 data have only 10 observations hence are not sufficient for the assessment. However, we concern about the appropriateness of asymptotic tests for small sample sizes, and examine this issue by simulation.

To evaluate the performance of the CZTP model for the finite sample sizes, we simulate the lesion data 10,000 times based on the CZTP model using (1) the original sample size, (2) the regression coefficients in Table 2 as true values, and (3) the original threshold value. Tables 1 and 2 (marked with †) provide the means and SE for counts and regression coefficients, respectively, and Figures 1 and 2 demonstrate the histograms of these estimates. The simulation of the DCIS.COM data shows minimal bias in the mean estimates while substantially large in several regression coefficients. The largest bias occurs in β_5 and β_8 for the groups treated with CD49^{hi}/CD24^{med} and CD44^{hi}/CD24^{med}, respectively, which have only 3 observations and low counts. The reason is, with small means (1.3 and 1.7 for these two groups), simulation can easily generate identical observations of 1, 1, and 1, which imply the truncated mean

$\hat{\mu} = \sum_{i=1}^n y_i / n = 1$. Solving Equation (2) for $\hat{\mu} = 1$ gives $\hat{\lambda} \approx 0$. According to Equation (4), consequently, $\hat{\beta}_0 + \hat{\beta}_5$ or $\hat{\beta}_0 + \hat{\beta}_8$ approaches negative infinity, thereby introduces downward bias in $\hat{\beta}_5$ or $\hat{\beta}_8$ since $\hat{\beta}_0$ is finite. For these cases, we added a small quantity 0.01 to the count values 1, which only alleviates but not eliminates bias. Figure 1 (b) shows an additional left peak (ranging between -6 and -4) in distributions of $\hat{\beta}_5$ and $\hat{\beta}_8$, which corresponds to the situation that all the 3 observations have the count 1 (or 1.01 with the adjustment) described above. Also

for this reason, the SEs of regression coefficients are much larger in simulation as compared to the SEs obtained from fitting the DCIS.COM data using the CZTP model (Table 2, columns CZTP model and Simulation†). Nonetheless, the SE of mean counts obtained from the data analysis and from simulation are quite similar (Table 1). On the other hand, when the mean is moderately large, e.g. 3.5 or greater, there is a small chance of obtaining observations all with count 1, and the aforementioned bias is less likely to happen (see $\widehat{\beta}_4$, $\widehat{\beta}_6$, and $\widehat{\beta}_7$ in Table 2 and Figure 1 (b)). We also examine the asymptotic property of estimators for regression coefficients using larger (10 times) sample sizes. Simulation (‡ in Table 2) indicates the distribution of estimates are quite symmetric (results not shown) and bias vanishes.

In contrast, the SUM-225 cell line has means larger than 5. It is unlikely to generate all one-count observations in any group in simulation, but censoring is more common (24% and 36% in the two groups, respectively). Among the 10,000 runs, 176 runs (< 2%) include one group or two that include all observations censored. These runs are discarded to avoid computational problems. Simulation based on the rest 9,824 runs confirms that the CZTP estimates, both the mean counts and regression coefficients, are accurate and precise.

We then examine Type I error rates by setting all regression coefficients except β_0 to zeros and conducting simulation using the same sample sizes compositions as the DCIS examples. Among the 10,000 simulation runs for the DCIS.COM experiment with mean count 2.6, the LR test rejects the null in 650 runs at 0.05 nominal level, which matches our understanding that LR test is slightly liberal. Using these 10,000 G^2 statistic values to form the hull distribution for the $G^2 = 14.2$ (from data analysis) to compare with, we correct the LR test p -value to be 0.095 from 0.077 under the $\chi^2(8)$ distribution. For the total of 43 observations composed of 3 to 9 per group, the Type I error rates range between 0.04 and 0.06 for testing $H_0: \mu_i = \mu_j$, $i \neq j$ by the Wald test. Nonetheless, the corresponding rates for the Wald test on $H_0: \beta_i = \beta_j$ range between 0.01 and 0.03, which are due to the “all 1-count observations” discussed above that create downward bias and inflated SE. These findings also confirm that comparison should be made on means μ rather than on regression coefficients β .

We also simulate 10,000 runs for the SUM-225 experiment but exclude 35 runs which contain all censored observations in one or both groups. For a total sample size of 10, (6 in one group plus 4 in the other) and the mean count 5.9, the Type I error rate of the LR test is 0.058. Comparing the $G^2 = 0.2$ from data analysis to the null G^2 distribution finds $p = 0.66$, quite similar to $p = 0.65$ determined by the $\chi^2(1)$ distribution shown above. The Wald test conservatively shows Type I error rate of 0.02 for testing both $H_0: \beta_1 = 0$ and $H_0: \mu_i = \mu_j$ and can undermine power. Thus, in this case, the LR test seems to be more reliable, which is consistent with literature (e.g. Agresti, 2002, p. 12). When sample sizes are large, say, 10 times as large as the examples, simulation results in 0.05 Type I error rate in both the LR and the Wald tests.

5. Discussion

The aim of this study is to assess the progression potential of human DCIS cells in animals. Experiments without positive growth were likely due to lack of progenitor cells and were excluded from analysis because they provided no information about progression potential. Only the fat pads with at least one mammary duct showing DCIS-like lesion were analyzed. Because some observations were right censored, we proposed a censored zero-truncated Poisson model for the analysis. Considering the small sample size, we also perform simulation to evaluate the performance of the proposed method. The simulation shows that the censored zero-truncated Poisson model in general provides reasonably accurate and precise estimates for analyzing positive count data with right censoring.

It should be noted that the parameters of interest are the *truncated* means whereas the regression coefficients in the log-linear model are linked to the *un-truncated* means. Although there is a one-to-one match between the truncated and un-truncated means, the difference exists between them but diminishes when the means increase (as shown in SUM-225 data). Therefore, inference in this content should be made on the truncated mean counts rather than on regression coefficients. The likelihood function based on the ZTP distribution allows comparing the truncated means through the likelihood ratio test which is more reliable than the Wald test as we expect for small sample sizes and moderate means.

The proposed CZTP model simultaneously considers zero-truncation and right censoring, thus both the ZTP model that handles left-truncation and the censored Poisson (CP, see Selvin, 1974, Terza, 1985, Caudill, 1995, and Jung et al., 2006) regression model that deals with right censoring can be viewed as two special cases of the CZTP model. When the mean count is small and censoring is not likely for reasonably large thresholds, the CZTP is virtually the same as ZTP model; when the mean counts are large such that the likelihood that zero count occurs is ignorable, the CZTP model converges to the CP model. However, if the mean count is moderately small, say, between 3 and 4, and larger counts are possible, censored observations are created due to a threshold such as the example of DCIS.COM cell line data, neither the ZTP nor the CP model is appropriate so we apply the CZTP model. The SUM-225 cell line, on the other hand, shows large mean counts of 5.9 and 6.7, thus the CP model can be a reasonable approximation. In fact, applying the CP model to the SUM-225 data produces the same results as shown in Tables 1 and 2.

Applying thresholds may have impacts on sample size in study design. For constant thresholds as in the DCIS example, censored proportion is a function of group means which in turn is a function of the un-truncated Poisson parameter. The greater the Poisson parameter (thus the mean and the variance) is, the more counts will be greater than the applied threshold and hence the larger the censored proportion becomes. Namely, applying thresholds lead to informative censoring and the censoring proportion of data can be derived from the probability distribution. Suppose one would like to compare two groups with truncated means of 2.50 and 3.44, the theoretical censored proportions are about 1% when threshold 7 is applied. On the other hand, the threshold of 7 causes theoretically 30% data censored for the two subgroups of SUM-225 with mean counts of 5.9 and 6.7. Figure 3 compares the power to detect the difference between the SUM-225 subgroups based on the CZTP model at 5% significance level for other thresholds of 9, 10, and 11 (19%, 11%, and 6% censored, respectively). If all data are recorded in their actual counts, a sample size of 306 or 153 per group is required for 80% power. Applying the thresholds of 10, 9, and 8 will lead to respective sample sizes of 312, 324, and 340. However, the threshold 11 does not increase the sample size. Such information provides statistical justification for applying thresholds and helps researchers to decide whether they would prefer to spending more time on extra experiments for additional animals due to introduction of thresholds, or to spending a few more minutes on obtaining the exact counts. We suggest researchers examining the impacts of applying thresholds using power analysis during the study design phase.

It is well known that count data are often over- and sometimes under-dispersed, and the models rely on Poisson distribution need modification. Several methods proposed to cope with this issue for standard Poisson model include adjusting the variance by a scale factor of the square root of the deviance or the Pearson's Chi-square statistic divided by the degrees of freedom, or applying the negative binomial (NB) instead of the Poisson model (Agresti, 2002). For positive count data, the ZTP model assumes the truncated variance less than the truncated mean (Equations (2) and (3)) and the difference diminish as the parameter of parent Poisson distribution increases. Hence, this method is not appropriate for over-dispersed data either. Should over dispersion be a concern, a zero-truncated negative binomial model can be applied.

See, for example, Lee et al. (2003) and Cruyff and van der Heijden (2008). Although censoring is not a concern in their work, the same method can be extended to incorporate censored observations. Alternatively, one can modify the censored generalized Poisson model proposed by Famoye and Wang (2004) to a zero-truncated version. For our DCIS example, we also fit the data with zero-truncated negative binomial model with censored observations taken into account. The dispersion parameter α ($\alpha > 0$ if overdispersed and a NB distribution converges to a Poisson when $\alpha \rightarrow 0$. See Equation (3) in Cruyff and van der Heijden, 2008) is estimated to be 0.45 ($p = 0.21$ from the Wald test, or $\chi^2_{(1)} = 1.9$, $p = 0.17$ from LR test).

In this study, we aim to evaluate the progression potential reflected by the positive counts of ducts showing DCIS-lesion while the original article of Behbod et al. targeted on the proportion of positive growth (positive counts). In a comprehensive analysis for larger sample sizes, it may be informative to jointly compare the proportions of positive growth and truncated means. To achieve this goal, one may apply the method analogous to the zero-inflated Poisson model except the Poisson component is replaced with the zero-truncated Poisson (with right censoring). Namely, let the count be a mixture of degenerated zeros with probability π_i and zero-truncated Poisson $f(y_i)$ with probability $1 - \pi_i$, then the probability function is given as $Pr(Y_i = y_i | \theta) = \mathbf{1}_{y_i=0} \times \pi_i + \mathbf{1}_{y_i>0} \times (1 - \pi_i)f(y_i)$ where π_i can be linked to covariates through a logit link. In other situations where combining zeros and positive counts are meaningful, excessive zeros often exist than expected and data are analyzed by the zero-inflated Poisson or a NB model which is a mixture of a Poisson or a NB distribution and a degenerate distribution of zero (see, e.g., Lambert 1992 and Yau et al., 2003).

In summary, the proposed CZTP model provides reasonably accurate and precise estimates for positive counts with small sample sizes and right censoring. Power analysis based on the CZTP model demonstrates the relationship between sample size and censoring threshold, which can be informative in study design, especially for experiments involve procedures that are expensive, low in success rates, and/or time/effort-consuming. Researchers who expect large counts and intend to apply thresholds should evaluate the impact of thresholds on sample size.

Acknowledgments

The authors thank the Associate Editor and the two anonymous referees for their very helpful comments and suggestion that greatly improve this paper. This research was sponsored by the National Institute of Health grant 1UL1RR033179-01.

APPENDIX

```
*****
**** This macro simulates zero-truncated Poisson count data with
**** right censoring at threshold h. The macro
**** (1) estimates the regression coefficients and truncated mean counts
**** (2) compares groups by reg. coeff. and by group means by
**** Likelihood ratio (LR) and Wald tests
**** (3) allows power analysis or examination of Type I error rate
**** for both LR and Wald tests
*****
;;
libname cztp "(your path to which simulated data will be saved)";
%macro dcis
(groups,b0,b1,b2,b3,b4,b5,b6,b7,b8,h,n0,n1,n2,n3,n4,n5,n6,n7,n8,alpha,simu);
/* groups = # of groups to be compared */
/* b0, b1, ..., b8 = regression coefficient in the log-linear model for
```



```

lambda, the un-truncated Poisson parameter */
/* h = the threshold value creating right censoring */
/* n0, n1, ..., n8 = the no. of observations in groups 0, 1, 2, ..., 8 */
/* alpha = the desired type I error rate */
/* simu = number of simulation runs */
%let g = %eval(&groups - 1);
data temp; keep set id g y x1 - x8;
n_cum1 = &n0 + &n1;
%do i = 2 %to 8; %let i_1 = %eval(&i - 1);
n_cum&i = n_cum&i_1 + &&n&i;
%end;
do set = 1 to &simu;
seed1 = 23456789 + set;
seed2 = 56789012 + set;
do id = 1 to n_cum8; /* generate a data with specified no. of groups and
no. observations per group by dummy variables */
x1 = 0; x2 = 0; x3 = 0; x4 = 0; x5 = 0; x6 = 0; x7 = 0; x8 = 0;
if id <= &n0 then do; g = 0; end;
if id > &n0 and id <= n_cum1 then do; x1 = 1; g = 1; end;
if id > n_cum1 and id <= n_cum2 then do; x2 = 1; g = 2; end;
if id > n_cum2 and id <= n_cum3 then do; x3 = 1; g = 3; end;
if id > n_cum3 and id <= n_cum4 then do; x4 = 1; g = 4; end;
if id > n_cum4 and id <= n_cum5 then do; x5 = 1; g = 5; end;
if id > n_cum5 and id <= n_cum6 then do; x6 = 1; g = 6; end;
if id > n_cum6 and id <= n_cum7 then do; x7 = 1; g = 7; end;
if id > n_cum7 and id <= n_cum8 then do; x8 = 1; g = 8; end;
lambda = exp(&b0 + &b1*x1 + &b2*x2 + &b3*x3 /* log-linear model for lambda */
+ &b4*x4 + &b5*x5 + &b6*x6 + &b7*x7 + &b8*x8);
u = ranuni(seed1);
t = -log(1 - u*(1 - exp(-lambda))); /* "time" to the first event */
t1 = lambda - t;
y = ranpoi(seed2, t1) + 1; /* the final zero-truncated count */
output;
end;
end;
run;
data temp2; set temp; h = &h;
if y >= h then do; z = h; censor = 1; end; /* create right censoring */
else do; z = y; censor = 0; end;
run;
proc sort data=temp2; by set g; run;
proc means data=temp2 noprint; by set g; /* compute the censored proportions
*/
var censor; output out=all_censor mean=p_censor;
run;
proc means data=temp2 noprint; by set g;
var z; output out=zavg mean=z_avg; /* compute means of truncated counts */
run;
data censor; keep set p_cens0 - p_cens&g;
merge %do il = 0 %to &g;
all_censor (where=(g=&il) rename=(p_censor=p_cens&il))
%end;;
run;

```

```

data poi0; merge temp2 censor; by set; run;
data poi; merge poi0 zavg; by set g;
if p_cens0 < 1 %do i2 = 0 %to &g; and p_cens&i2 < 1 %end;;
/* exclude a simulation run if any group has all obs. censored */
if z_avg = 1 then do;
z = 1 + 0.01; /* add 0.01 to observations where all
observations have count 1 in a group */
end;
run;
title "CZTP: censored at h=&h";
proc nlmixed data=poi; by set;
parms %do j1 = 0 %to &g; b&j1 = 0 %end;;
lambda = exp(b0 %do j2 = 1 %to &g; + b&j2*x&j2 %end;);
if censor = 0 then
ll = z*log(lambda) - lgamma(z + 1) - log(exp(lambda) - 1); /* log-likelihood
for un-censored obs */
else if censor = 1 then do; pr_cum = 0;
do i = 1 to h - 1;
pr_cum + lambda**i/gamma(i + 1)/(exp(lambda) - 1); /* log-likelihood for
censored obs */
end;
ll = log(1 - pr_cum);
end;
model z ~ general(ll);
estimate "mu0" exp(b0)/(1 - exp(-exp(b0))); /* mean count of group 0 */
%do j3 = 1 %to &g;
estimate "mu&j3" exp(b0 + b&j3)/(1 - exp(-exp(b0 + b&j3))); /* mean count of
group j3 */
estimate "mu0 vs. mu&j3" /* difference in mean counts b/w groups 0 & j3 */
exp(b0)/(1-exp(-exp(b0))) - exp(b0+b&j3)/(1-exp(-exp(b0 + b&j3)));
%end;
%do k = 2 %to &g; %let k_1 = %eval(&k - 1);
%do k2 = 1 %to &k_1;
estimate "b&k2 vs. b&k" b&k2 - b&k; /* difference in reg. coeff. b/w groups
*/
estimate "mu&k2 vs. mu&k" /* difference in mean counts b/w groups */
exp(b0 + b&k2)/(1-exp(-exp(b0+b&k2)))
- exp(b0 + b&k)/(1 - exp(-exp(b0 + b&k)));
%end;
%end;
ods output FitStatistics=fit; /* max. log-likelihood */
ods output ParameterEstimates=parms; /* regression coefficients */
ods output AdditionalEstimates=parms_a; /* mean counts and b/w-group
comparisons */
run;
title "CZTP: H0";
proc nlmixed data=poi; by set;
parms b0 = 0; lambda = exp(b0);
if censor = 0 then
ll = z*log(lambda) - lgamma(z + 1) - log(exp(lambda) - 1);
else if censor = 1 then do; pr_cum = 0;
do i = 1 to h - 1;
pr_cum + lambda**i/gamma(i + 1)/(exp(lambda) - 1);

```

```

end;
ll = log(1 - pr_cum);
end;
model z ~ general(ll);
ods output FitStatistics=fit0; /* max. log-likelihood under H0*/
run;
data LL; merge /* Likelihood ratio test */
fit (where=(descr="-2 Log Likelihood") rename=(value=_2LL))
fit0 (where=(descr="-2 Log Likelihood") rename=(value=_2LL0));
by set;
Chisq = _2LL0 - _2LL; df = &g; pval = 1 - probchi(Chisq,df);
if pval < &alpha then reject = 1; else reject = 0;
drop descr;
run;
data parms_a; set parms_a;
if probt < &alpha then reject = 1;
else reject = 0;
run;
data CZTP.dcis_poi&simu._runs; set poi; run;
data CZTP.dcis&simu._runs; set parms; run;
data CZTP.dcis_a&simu._runs; set parms_a; run;
data CZTP.dcis_ll&simu._runs; set ll; run;
%dif;
%mend;
%macro dif;
title "Power/Type I error rate of LRT";
proc means data=cztp.dcis_ll&simu._runs n mean; var reject; run;
data dif_mu; set cztp.dcis_a&simu._runs;
where label = "mu0"
%do i = 1 %to &g; %let i_1 = %eval(&i - 1);
%do i2 = 0 %to &i_1;
or label = " mu&i2 vs. mu&i"
%end;
%end;;
run;
proc sort data=dif_mu; by label; run;
title "Power/Type I error rate of differences in mu (LR test)";
proc means data=dif_mu n mean; var reject; by label; run;
data b; set cztp.dcis&simu._runs;
if probt < &alpha then reject=1; else reject = 0;
run;
proc sort data=b; by parameter; run;
data dif_b; set cztp.dcis_a&simu._runs;
where label = "b0"
%do i = 2 %to &g; %let i_1 = %eval(&i - 1);
%do i2 = 1 %to &i_1;
or label = "b&i2 vs. b&i"
%end;
%end;;
run;
proc sort data=dif_b; by label; run;
title "Power/Type I error rate of diff in b (LR test)";
proc means data=b n mean; var reject; by parameter; run;

```

```
proc means data=dif_b n mean; var reject; by label; run;
%mend;
%dcis(9,0.8513,0.2288,-0.1988,-0.3853,0.4442,-1.3524,0.3987,0.4336,
-0.7324,8,9,4,4,4,4,3,5,7,3,0.05,10000);
```

References

- Agresti, AA. Categorical Data Analysis. 2nd edition. John Wiley and Sons, Inc.; 2002.
- Allred DC, Wu Y, Mao S, Nagtegaal ID, Lee S, Perou CM, Mohsin SK, O'Connell P, Tsimelzon A, Medina D. Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. *Clinical Cancer Research*. 2008; 14(2):370–378. [PubMed: 18223211]
- Behbod F, Kittrell FS, LaMarca H, Edwards D, Kerbawy S, Heestand JC, Young E, Mukhopadhyay P, Yeh HW, Allred DC, Hu M, Polyak K, Rosen JM, Medina D. An intra-ductal human-in-mouse transplantation model mimics the subtypes of ductal carcinoma in situ. *Breast Cancer Research*. 2009; 11(5):R66. DOI: 10.1186/bcr1847. [PubMed: 19735549]
- Boughey JC, Gonzalez RJ, Bonner E, Kuerer HM. Current treatment and clinical trial developments for ductal carcinoma in situ of the breast. *Oncologist*. 2007; 12(11):1276–87. [PubMed: 18055847]
- Caudill SB. Modeling household fertility decisions: Estimation and testing of censored regression models for count data. *Empirical Economics*. 1995; 20:183–196. [PubMed: 12320575]
- Cohen AC. Estimation of the Poisson parameter from truncated samples and from censored samples. *Journal of the American Statistical Association*. 1954; 49(265):158–168.
- Cruyff MJLF, van der Heijden PGM. Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model. *Biometrical Journal*. 2008; 50(6):1035–1050. [PubMed: 19067336]
- Famoye F, Wang W. Censored generalized Poisson regression model. *Computational Statistics and Data Analysis*. 2004; 46:547–560.
- Fokianos K. Truncated Poisson Regression for time series of counts. *Scandinavian Journal of Statistics*. 2001; 28:645–659.
- Grogger JT, Carson RT. Models for truncated counts. *Journal of Applied Econometrics*. 1991; 6(3): 225–238.
- Gurmu S. Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business & Economic Statistics*. 1991; 9(2):215–222.
- Jung YC, Jhun M, Song SH. Testing for overdispersion in a censored Poisson regression model. *Statistics*. 2006; 40(6):533–543.
- Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992; 34(1):1–14.
- Lee AH, Wang K, Yau KKW, Somerford PJ. Truncated negative binomial mixed regression modeling of ischaemic stroke hospitalizations. *Statistics in Medicine*. 2003; 22(7):1129–1139. [PubMed: 12652558]
- Matthews JN, Appleton DR. An application of the truncated Poisson distribution to immunogold assay. *Biometrics*. 1993; 49(2):617–621. [PubMed: 8369393]
- Puza BD, Johnson HL, O'Neill TJ. Bayesian truncated Poisson regression with application to Dutch illegal immigration data. *Communications in Statistics - Simulation and Computation*. 2008; 37(8): 1565–1577.
- [R] simulated zero-truncated Poisson distribution. <https://stat.ethz.ch/pipermail/r-help/2005-May/070678.html>
- Saei A, Chambers R. Modelling truncated and clustered count data. *Australian & New Zealand Journal of Statistics*. 2005; 47(3):339–349.
- Selvin S. Maximum Likelihood Estimation in the truncated or censored Poisson distribution. *Journal of the American Statistical Association*. 1974; 69(345):234–237.
- Shackleton M, Vaillant F, Simpson KJ, Stingl J, Smyth GK, Asselin-Labat ML, Wu L, Lindeman GJ, Visvader JE. Generation of a functional mammary gland from a single stem cell. *Nature*. 2006; 439 (7072):84–88. [PubMed: 16397499]

- Silvia JMCS. Generalized Poisson regression for positive count data. *Communication in Statistics – Simulation and Computation*. 1997; 26(3):1089–1102.
- Terza JV. A Tobit-type estimator for the censored Poisson regression model. *Economics Letters*. 1985; 18:361–365.
- van der Heijden PGM. Estimating the size of a criminal population from police records using the truncated Poisson Regression Model. *Statistica Neerlandica*. 2003; 57(3):289–304.
- Xie T, Aickin M. A truncated Poisson regression model with applications to occurrence of adenomatous polyps. *Statistics in Medicine*. 1997; 16(16):1845–1857. [PubMed: 9280037]
- Yau KW, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of overdispersed count data with extra zeros. *Biometrical Journal*. 2003; 45(4):437–452.

Figure 1. Simulated distributions of (a) the truncated mean counts $\dagger$ and (b) regression coefficients for DCIS.COM

$$\dagger \hat{\mu}_k = \hat{\lambda}_k / (1 - e^{-\hat{\lambda}_k}) \text{ where } \hat{\lambda}_0 = \exp(\hat{\beta}_0) \text{ and } \hat{\lambda}_\ell = \exp(\hat{\beta}_0 + \hat{\beta}_\ell), \ell \neq 0$$

Figure 2. Simulated distributions of (a) the meancounts $\dagger$ and (b) regression coefficients for SUM-225
 $\dagger \hat{\mu}_k = \hat{\lambda}_k / (1 - e^{-\hat{\lambda}_k})$ where $\hat{\lambda}_0 = \exp(\hat{\beta}_0)$ and $\hat{\lambda}_1 = \exp(\hat{\beta}_0 + \hat{\beta}_1)$

Figure 3. Impacts of thresholds on power and sample size when CZTP model is applied for $\mu_1 = 5.9$, $\mu_2 = 6.7$.

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

The mean counts and SE for number of mammary ducts with DCIS-like lesions

Cell line	Group	# injections	# of ducts positive	Truncated mean counts (SE)	
				CZTP model [†]	Simulation [‡]
DCIS.COM	G0. Random sort	28	1, 1, 1, 2, 2, 3, 4, 7+	2.6 (0.47)	2.6 (0.47)
DCIS.COM	G1. CD49 ^{thi} /MUC-1 ^{med}	23	1, 1, 2, 7+	3.1 (0.83)	3.1 (0.81)
DCIS.COM	G2. CD49 ^{thi} /MUC-1 ^{lo}	6	1, 2, 2, 4	2.3 (0.61)	2.2 (0.61)
DCIS.COM	G3. CD49 ^{med} /MUC-1 ^{lo}	23	1, 2, 2, 3	2.0 (0.54)	2.0 (0.54)
DCIS.COM	G4. CD49 ^{med} /MUC-1 ^{med}	6	1, 2, 6, 6	3.8 (0.92)	3.8 (0.94)
DCIS.COM	G5. CD49 ^{thi} /CD24 ^{med}	8	1, 1, 2	1.3 (0.35)	1.3 (0.35)
DCIS.COM	G6. CD49 ^{thi} /CD24 ^{lo}	8	2, 3, 3, 5, 5	3.6 (0.80)	3.6 (0.81)
DCIS.COM	G7. CD44 ^{hi} /CD24 ^{lo}	24	1, 3, 3, 4, 4, 4, 7	3.7 (0.69)	3.7 (0.70)
DCIS.COM	G8. CD44 ^{hi} /CD24 ^{med}	6	1, 1, 3	1.7 (0.51)	1.7 (0.51)
SUM-225	G0. CD49 ^{med} /MUC-1 ^{hi}	11	2, 5, 5, 7+, 7+	5.9 (1.04)	6.0 (1.09)
SUM-225	G1. CD49 ^{med} /CD24 ^{hi}	6	2, 6, 7+, 7+	6.7 (1.44)	6.9 (1.43)

[†] Mean and SE obtained from fitting the DCIS data using the proposed method

[‡] Based on 10,000 runs of simulation using the estimates from the fourth column (CZTP model) in Table 2 as the true values and the sample size as shown in the DCIS example

\$watermark-text

\$watermark-text

\$watermark-text

Table 2

Regression coefficient estimates and SE

Cell line	Group	Parameter in Eq.(4)	Mean (SE)	
			CZTP model [†]	Simulation [‡]
DCIS.COM	G0. Random sort	β_0	0.85 (0.24)	0.81 (0.26)
DCIS.COM	G1. CD49 ^{thi} /MUC-1 ^{med}	β_1	0.23 (0.40)	0.20 (0.48)
DCIS.COM	G2. CD49 ^{thi} /MUC-1 ^{lo}	β_2	-0.20 (0.47)	-0.32 (0.71)
DCIS.COM	G3. CD49 ^{med} /MUC-1 ^{lo}	β_3	-0.39 (0.52)	-0.58 (0.90)
DCIS.COM	G4. CD49 ^{med} /MUC-1 ^{med}	β_4	0.44 (0.36)	0.44 (0.40)
DCIS.COM	G5. CD49 ^{thi} /CD24 ^{med}	β_5	-1.35 (0.99)	-2.42 (1.87)
DCIS.COM	G6. CD49 ^{thi} /CD24 ^{lo}	β_6	0.40 (0.35)	0.41 (0.38)
DCIS.COM	G7. CD44 ^{hi} /CD24 ^{lo}	β_7	0.43 (0.32)	0.45 (0.34)
DCIS.COM	G8. CD44 ^{hi} /CD24 ^{med}	β_8	-0.73 (0.70)	-1.34 (1.57)
SUM-225	G0. CD44 ^{med} /CD24 ^{med}	β_0	1.77 (0.18)	1.76 (0.19)
SUM-225	G1. CD49 ^{med} /MUC-1 ^{hi}	β_1	0.14 (0.28)	0.14 (0.28)

[†] Mean and SE obtained from fitting the DCIS data using the proposed method

[‡] Based on 10,000 runs of simulation using the estimates from the fourth column (CZTP model) as the true values and the sample size as shown in the DCIS example

[§] Increasing the sample sizes by 10 times