Finding the right distribution for highly skewed zero-inflated clinical data

Resmi Gupta[1], Bradley S. Marino[1], James F. Cnota[1], Richard F. Ittenbach[1]

ABSTRACT

Discrete, highly skewed distributions with excess numbers of zeros often result in biased estimates and misleading inferences if the zeros are not properly addressed. A clinical example of children with electrophysiologic disorders in which many of the children are treated without surgery is provided. The purpose of the current study was to identify the optimal modeling strategy for highly skewed, zero-inflated data often observed in the clinical setting by: (a) simulating skewed, zero-inflated count data; (b) fitting simulated data with Poisson, Negative Binomial, Zero-Inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) models; and, (c) applying the aforementioned models to actual, highly skewed, clinical data of children with an EP disorder. The ZIP model was observed to be the optimal model based on traditional fit statistics as well as estimates of bias, mean-squared error, and coverage.

Key words: Zero-inflated Poisson; Skewed data; Generalized linear model

(1) Division of Biostatistics and Epidemiology and The Heart Institute, Cincinnati Children's Hospital Medical Center (USA)

CORRESPONDING AUTHOR: *Resmi Gupta, Division of Biostatistics and Epidemiology, MLC 5041, 3333 Burnet Avenue, Cincinnati, OH 45229 (USA).*
e-mail: resmi.gupta@cchmc.org.
DOI: ***10.2427/8732***

INTRODUCTION

Modeling discrete data in the health sciences continues to pose a challenge even for the most experienced researchers. For discrete outcomes, common methods of data analysis typically involve Poisson and negative binomial modeling strategies. However, these seemingly simple and straightforward approaches to modeling may not be appropriate when observations include large numbers of zeros. Researchers must then consider a new class of models that provides a more flexible way to address the discrete data with large numbers of zeros in the dependent variable [1].

Data obtained in clinic settings very often yield distributions that are anything but normal. For example, in pediatric cardiovascular research, length of hospital stay (in total or in the intensive care unit), number of outpatient visits during a given period of time, and children presenting with a certain condition, with and without surgery, offer examples in this regard. Given the characteristics of many of these distributions, simply using traditional parametric techniques which fail to meet the assumptions is likely to produce misleading inferences and conclusions. As a result, zero-inflated

models appear to be gaining favor amongst statisticians. For example, Bohning et al. [2] have applied zero-inflated models when evaluating intervention effects for decayed, missing and filled teeth in dental epidemiology, while Cheung [3] has reported using them in inquiries involving early growth and development studies.

This manuscript uses data from pediatric cardiovascular clinical care to illustrate the problem. In order to maximize the accuracy and utility of researchers' findings for modeling the number of surgeries for children with electrophysiological (EP) disorders, the use of zero-inflated models may be a more appropriate way to avoid bias and a more efficient way of fitting models. To assess the effects of covariates on the aforementioned outcome, zero-inflated models will be used to estimate parameters and help accommodate violations of underlying model assumptions.

The purpose of this paper is to identify the optimal modeling strategy for highly skewed, discrete clinical data with excess numbers of zeros using the following three-step approach. First, by simulating a distribution of skewed, zero-inflated count data; second, by fitting simulated data with Poisson, Negative Binomial, Zero-Inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) models; and third, by applying the aforementioned models to actual, highly skewed, clinical data of children with an EP disorder.

METHODS

Study population

The records of 286 children ranging from 8 to 18 years of age at the time of their outpatient cardiology visit were used for the clinical portion of this study. All children in the sample were diagnosed with an EP disorder, a form of cardiac disease that primarily affects heart rhythm. These non-transplant EP children had structurally normal hearts but have undergone heart surgery in the form of a catheter-based intervention or implantation of a pacemaker. This study population consisted of 147 males and 139 females.

Clinical variables

Data from four variables were used to model the clinic data: cardiac related hospitalization, heart block, prematurity, and time. *Heart block* was characterized as the presence or absence of a heart block. Children with heart block had either second-degree atrioventricular block (type 1 or type 2), or congenital or acquired form of complete heart block. *Cardiac related hospitalization* was defined as a binary variable (0 = 2 or fewer visits per year; 1 = more than 2 visits per year). *Prematurity* was defined as any child born prior to 37 weeks gestation. *Time* was defined simply as "time in months" since last hospitalization. Except for time, which was treated as a continuous variable, the other three variables were treated as binary variables. This study was a part of a larger study for which the approval was obtained from the Institutional Review Board at Cincinnati Children's Hospital Medical Center.

Statistical Analyses

Data analysis proceeded in three discrete stages: (a) *simulation phase*, which involved simulating distributions of highly skewed data that would mimic data observed in the clinical setting; (b) *modeling phase,* which consisted of modeling the simulated data using four different regression-based techniques (Poisson, Negative Binomial, Zero-inflated Poisson and Zero-inflated Negative Binomial); and (c) *application phase,* which involved modeling actual clinical data using the same four regression-based techniques and evaluating model performance in an effort to identify the optimal analytical technique for clinicians and applied biomedical researchers. All data were analyzed using SAS v9.2.

Simulation Phase

Data were generated within a mixture distribution framework involving two-step process. In order to estimate the zero-part and positive part of the distribution separately, a binomial (n, p_0) distribution was used to estimate the probability of zeros in the data, and a poisson (μ) distribution was used to estimate the count part of the data in the simulation process. To determine the estimates for the zero part, the probability (p_0) was generated using a logistic distribution based on zero-inflated covariates. The conditional mean of the Poisson distribution was generated by fitting a Poisson distribution based on the covariates from the positive count portion of the distribution. To reflect the data, 3 covariates, an intercept, a categorical surrogate and a continuous surrogate variable were used for both parts of the mixture model. The true values of the estimable parameters for both parts of the mixture distribution were based on the estimates obtained from real data. While the categorical covariate was used to mimic heart block (Y/N), the continuous covariate was used to mimic time since last hospitalization for children with EP disorders. A total of three sets of simulations were conducted with n=1 000 000 for each set.

Modeling Phase

In the modeling phase, the simulated data were fit using Poisson, Negative Binomial, Zero-inflated Poisson, and Zero-inflated Negative Binomial probability distributions (see appendix A for the probability distribution functions). Model performance and estimates of precision for each of the aforementioned models were calculated using bias, mean squared error (MSE) and coverage probability [4].

Application Phase

In the application phase, analyses proceeded in a two-step sequence: first, descriptive statistics were computed to describe the basic features of the data and second, the four models discussed previously were then applied to the clinical data. A Poissonness plot [5] was generated to determine if the data were likely to have come from a Poisson distribution while a Lagrange multiplier (LM) [6, 7] test was used to check for model over-dispersion. Introduced by Hoaglin, a Poissonness plot is a graphical measure based on the Poisson distribution. If the data came from a Poisson distribution, then a plot of the sum of the natural logarithm of a frequency of y, and the natural logarithm of $y!$ against y should form a straight line. See appendix B for the mathematical derivation of the Poissonness plot. Cameron and Trivedi proposed a LM statistic for overdispersion in the Poisson model versus the negative binomial regression model. Under the null hypothesis of the Poisson model with no overdispersion, the limiting distribution of the LM statistic would follow a chi-square distribution with 1 degree of freedom. See appendix C for the functional form of the LM statistic. Van den Broek score tests [8] were used to formally test for zero inflation in the data. The Van den Broek statistic is based on a comparison of actual zeros to those predicted by the model to test for zero-inflation relative to a Poisson distribution. Under the null hypothesis of no zero-inflation, the test follows a chi-square distribution with 1 degree of freedom. See appendix D for the score test formula.

Each of the four models was compared using log-likelihood estimates as a measure of model performance. In addition, the Vuong test statistic [9] was used to compare non-nested models (e.g. Poisson versus Zero-inflated Poisson). The Vuong test (see appendix E) is based on a comparison of the predicted probabilities of two non-nested models. We examined model fit by comparing the Akaike information criterion [10] (AIC) and predicted and observed probabilities of each count outcome for each probability distribution individually.

RESULTS

Simulation Phase

The purpose of the simulation phase of the study was to generate samples of data similar to

that frequently observed in the clinical setting, specifically, skewed data with large numbers of zeros. The initial set of true values for the intercept and covariates were based on a Pediatric Cardiac Quality of Life Inventory data set in order to represent as closely as possible data obtained in clinical practice. In the first set of simulations, six parameters were estimated by using an intercept (α_{00}=1.56), covariate 1 (α_{01}= -1.37) to mimic heart block(s), and covariate 2 (α_{02}=-0.04) to mimic time since last hospitalization for the *zero portion* of the distribution, and then an intercept (α_{10}=0.21), covariate 1 (α_{11}=0.72) and covariate 2 (α_{12}= - 0.02) for the *zero portion* of the distribution (Table 1). To increase the generalizability of the simulation results, two additional sets of simulations were conducted for the same six parameters. True values for the second and third sets of simulations were derived by adding increments of 0.10 to the true values of the previous sets for intercepts ($\alpha_{00,}$ α_{10}) and covariate 1 ($\alpha_{01,}$ α_{11}), and by adding 0.01 to the true values of the previous sets for covariate 2 ($\alpha_{02,}$ α_{12}). See Table 1 for a list of true values used in this simulation.

Modeling phase for the simulated data

Once the data were simulated, four different modeling strategies were used to evaluate model fit: Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial. Zero-inflated Poisson and zero-inflated negative binomial distributions had binary and continuous covariates

TABLE 1

"TRUE" VALUES USED TO GENERATE SIMULATED DATA

Parameter	True Value
First Simulation	
Zeros	
Intercept (α_{00})	1.56
Covariate 1 (α_{01})	-1.37
Covariate 2 (α_{02})	-0.04
Positives	
Intercept (α_{10})	0.21
Covariate 1 (α_{11})	0.72
Covariate 2 (α_{12})	-0.02
Second Simulation	
Zeros	
Intercept (α_{00})	1.66
Covariate 1 (α_{01})	-1.27
Covariate 2 (α_{02})	-0.03
Positives	
Intercept (α_{10})	0.31
Covariate 1 (α_{11})	0.82
Covariate 2 (α_{12})	-0.01
Third Simulation	
Zeros	
Intercept (α_{00})	1.76
Covariate 1 (α_{01})	-1.17
Covariate 2 (α_{02})	-0.02
Positives	
Intercept (α_{10})	0.41
Covariate 1 (α_{11})	0.92
Covariate 2 (α_{12})	-0.009

along with intercept terms for both the zero and positive portions of the model. Table 2 reflects the estimated mean and empirical standard error for each parameter of interest in the three sets of simulations. The empirical standard error and the average of the within simulation standard errors were almost the same for zero-inflated Poisson and zero-inflated negative binomial models, which is not the case for standard Poisson or negative binomial models. Consequently, both Poisson and negative binomial models would be problematic in making accurate inferences about the parameters as well as obtaining correct coverage.

Table 3 reflects the criteria for evaluating the performances of each modeling technique in terms of comparison between the simulated results and true values used to simulate the data for three sets of simulation. Estimates of bias, mean squared error and coverage probabilities are reported in Table 3. For each coefficient, bias appeared to be much larger for both Poisson and negative binomial models compared to the zero-inflated Poisson and zero-inflated negative binomial models in all three sets of simulations. This finding suggests greater accuracy for zero-inflated Poisson and zero-inflated negative binomial models compared to standard Poisson and negative binomial models. For all three sets of simulations, MSEs were much smaller for zero-inflated Poisson and zero-inflated negative binomial models compared to standard Poisson and negative binomial models. The criterion for acceptability of coverage was set such that the coverage should fall within two standard errors of the nominal coverage probability [4]. The average length of the 95% confidence interval for the parameter estimate $(\hat{\delta})$ was considered as a tool for assessment of coverage. Both zero-inflated Poisson and zero-inflated negative binomial models maintained correct coverage probability compared to the Poisson and negative binomial models.

In summary, these results suggest that for zero-inflated discrete distributions, fitting with a standard Poisson or negative binomial model are likely to lead to misleading inferences about parameters. The increased accuracy and precision for the zero-inflated Poisson and zero-inflated negative binomial models suggest improvements in efficiency and power compared to standard Poisson or negative binomial models.

Application to Clinical Data

The records of 286 non-transplanted children with EP disorders were used for this portion of the study. Of this number, 87% (n = 249) had no reported surgical procedures, resulting in an unusually high number of zeros for the current sample. See Figure 1 for an illustration of number of surgeries. The mean age for this population is 13.2 years (SD = 3.0 years). The maximum number of surgeries reported in the sample was five with 6.3% (n = 18) of children having one surgery and 3.85% (n = 11) of children reporting 2 surgeries. Children with 3, 4 or 5 surgeries accounted for 1.4% (n = 4), 0.35% (n = 1) and 0.35 %(n = 1) of the data, respectively. Roughly 70% of the children had no more than two cardiac related hospitalizations (n = 202). Moreover, in this non-transplant EP population, children born prematurely accounted for 15.7% of the sample (n = 45). Approximately 18.5% of the children included in the study had heart block (n = 53) and 18 children (34%) with heart block were born prematurely. The average time since last hospitalization was 45.19 months (SD = 46.62 months).

With respect to model diagnostics, a Poissonness plot [5] was used to determine graphically whether the Poisson distribution was an appropriate model for this sample.

The clear curvature of this relationship (Figure 2) suggested that the Poisson distribution did not provide a good fit to these data. The Lagrange multiplier (LM) test [6, 7] statistic was 4.48 (p= 0.03), suggesting the existence of overdispersion due to heterogeneity in the sample and violating the Poisson assumption of equidispersion. The Van den Broek score test statistic [8] was used and yielded a statistically significant result suggesting existence of overdispersion due to extra zeros [11, 12, 13].

With respect to model performance, the log-likelihood (LL) was used as a measure of each model's performance. Table 4 clearly shows an improvement in model fitting from Poisson (LL = -108.64) and negative binomial (LL = -105.10) to zero-inflated Poisson (LL = -90.63), and zero-inflated negative binomial (LL = -90.65) models. The Vuong test statistic [9] result reflected that zero-inflated Poisson performed better than standard Poisson (3.54, p<0.01), which also holds for zero-inflated negative binomial vs. negative binomial (3.85, p<0.01).

TABLE 2

SIMULATION RESULTS FOR EACH SET OF "TRUE" VALUES

Parameter	True Value	Poisson		NB		ZIP		ZINB	
		M (SD)	M (eSE)	M (SD)	M (eSE)	M (SD)	M (eSE)	M (SD)	M (eSE)
First Simulation									
Zeros									
Intercept (α_{00})	1.56					1.55 (0.17)	0.17	1.55 (0.17)	0.17
Covariate 1 (α_{01})	-1.37					-1.37 (0.20)	0.19	-1.37 (0.20)	0.20
Covariate 2 (α_{02})	-0.04					-0.04 (0.08)	0.08	-0.04 (0.08)	0.08
Positives									
Intercept (α_{10})	0.21	-1.56 (0.14)	0.09 -1.56 (0.14)	0.12	0.19 (0.13)	0.14	0.18 (0.13)	0.13	
Covariate 1 (α_{11})	0.72	1.69 (0.15)	0.10 1.70 (0.15)	0.14	0.73 (0.14)	0.14	0.73 (0.14)	0.13	
Covariate 2 (α_{12})	-0.02	0.003 (0.06)	0.03 0.004 (0.06)	0.04	-0.02 (0.04)	0.04	-0.02 (0.04)	0.04	
Second Simulation									
Zeros									
Intercept (α_{00})	1.66					1.66 (0.16)	0.16	1.66 (0.16)	0.16
Covariate 1 (α_{01})	-1.27					-1.27 (0.19)	0.19	-1.27 (0.19)	0.19
Covariate 2 (α_{02})	-0.03					-0.03 (0.08)	0.08	-0.03 (0.08)	0.08
Positives									
Intercept (α_{10})	0.31	-1.54(0.14)	0.09 -1.54 (0.14)	0.12	0.30 (013)	0.13	0.29 (0.13)	0.12	
Covariate 1 (α_{11})	0.82	1.76(0.16)	0.10 1.76 (0.16)	0.14	0.82 (0.14)	0.14	0.82 (0.14)	0.13	
Covariate 2 (α_{12})	-0.01	0.01(0.06)	0.03 0.01 (0.06)	0.05	-0.01 (0.04)	0.04	-0.01 (0.04)	0.04	
Third Simulation									
Zeros									
Intercept (α_{00})	1.76					1.76 (0.16)	0.16	1.76 (0.16)	0.16
Covariate 1 (α_{01})	-1.17					-1.17 (0.18)	0.18	-1.17 (0.18)	0.18
Covariate 2 (α_{02})	-0.02					-0.02 (0.08)	0.08	-0.02 (0.08)	0.08
Positives									
Intercept (α_{10})	0.41	-1.52(0.14)	0.09 -1.53 (0.14)	0.13	0.40 (0.13)	0.13	0.39 (0.13)	0.12	
Covariate 1 (α_{11})	0.92	1.82(0.16)	0.10 1.82 (0.16)	0.18	0.93 (0.14)	0.14	0.93 (0.14)	0.13	
Covariate 2 (α_{12})	-0.009	0.006(0.06)	0.03 0.006 (0.07)	0.08	-0.008 (0.03)	0.03	-0.008 (0.03)	0.03	

Note: NB, Negative Binomial; ZIP, Zero-inflated Poisson; ZINB, Zero-inflated Negative Binomial; M(SD), Mean(Standard Deviation); M(eSE), Mean of Estimated Standard Error

TABLE 3

ACCURACY AND PRECISION OF SIMULATION RESULTS FOR EACH SET OF TRUE VALUES FOR THE POSITIVE PART OF THE DISTRIBUTION FITTED BY FOUR MODELS

Distribution	Intercept (α_{10}) B	MSE	C	Covariate 1 (α_{11}) B	MSE	C	Covariate 2 (α_{12}) B	MSE	C
First Simulation									
Poisson	-1.77	3.15	0.00	0.97	0.96	0.00	0.02	0.004	0.56
NB	-1.77	3.15	0.00	0.98	0.98	0.00	0.02	0.004	0.57
ZIP	-0.02	0.01	0.96	0.01	0.01	0.96	0.00	0.001	0.95
ZINB	-0.03	0.01	0.95	0.01	0.01	0.94	0.00	0.001	0.94
Second Simulation									
Poisson	-1.85	3.44	0.00	0.94	0.90	0.00	0.02	0.004	0.50
NB	- 1.85	3.44	0.00	0.94	0.90	0.00	0.02	0.004	0.50
ZIP	-0.01	0.01	0.95	0.00	0.01	0.96	0.00	0.001	0.95
ZINB	-0 02	0.01	0.95	0.00	0.01	0.94	0.00	0.001	0.95
Third Simulation									
Poisson	-1.93	3.74	0.00	0.90	0.83	0.00	0.01	0.003	0.45
NB	-1.94	3.78	0.00	0.90	0.83	0.00	0.01	0.005	0.46
ZIP	-0.01	0.01	0.95	0.01	0.01	0.95	0.001	0.0009	0.94
ZINB	-0.02	0.01	0.95	0.01	0.01	0.94	0.001	0.0009	0.95

Note: NB, Negative Binomial; ZIP, Zero-inflated Poisson; ZINB, Zero-inflated Negative Binomial; B, Bias; MSE, Mean Squared Error; C, Coverage Probability

TABLE 4

CLINICAL DATA: MODEL SELECTION

Test statistic (*p-value*)	Poisson	NB	ZIP	ZINB	Vuong test
Log-likelihood	-108.64	-105.10	-90.63	-90.65	
AIC	231.28	220.20	205.27	205.28	
Vuong test (Poisson vs. ZIP)					3.54 (<0.01)
Vuong test (NB vs. ZINB)					3.85 (<0.01)
Estimated proportion of zeros (%)	0.771	0.783	0.875	0.870	
Estimated dispersion Parameter	NA	0.67	NA	0.001	

Note: NB, Negative Binomial; ZIP, Zero-inflated Poisson; ZINB, Zero-inflated Negative Binomial

TABLE 5

CLINICAL DATA: RESULTS FROM ZERO-INFLATED POISSON MODEL

Risk Factor	β Coefficient	SE	*t*-value	*p*-value
Logistic Portion of the Model				
Intercept	0.01	2.14	0.01	0.99
Hospitalization	2.25	2.33	0.97	0.33
Heart block	0.12	1.27	0.10	0.92
Prematurity	6.99	3.71	1.88	0.05
Time	-0.08	0.03	-2.48	0.01
Heart block x prematurity	-7.49	3.91	-1.91	0.05
Poisson Portion of the Model				
Intercept	-2.19	0.66	-3.32	< 0.01
Hospitalization	1.81	0.53	3.40	< 0.01
Heart block	1.48	0.52	2.87	< 0.01
Prematurity	2.65	1.04	2.54	0.01
Time	-0.01	0.01	-2.32	0.02
Heart block x prematurity	-2.26	1.13	-2.01	0.04

Note: The logistic portion of the table provides results for the portion of the data that consist of always zero while the Poisson portion of the table consists of the sampling zeroes as well as positive integer portion of the data

Next, predicted and observed probabilities of each outcome were compared. The AIC values for zero-inflated Poisson (AIC=205.27) and zero-inflated negative binomial (AIC=205.28) models were smaller compared to standard Poisson (AIC=231.28) and standard negative binomial (AIC=220.20) models and, hence, suggested better fit for the data using zero-inflated regression. The zero-inflated Poisson and zero-inflated negative binomial models predicted each count outcome very close to the observed counts, suggesting better fit than standard Poisson and negative binomial models. Based on the above mentioned criteria for model selection and evaluation, we opted for the zero-inflated Poisson model for fitting the clinical data.

Count portion of the model

In the count (Poisson) portion of the model, hospitalization, presence of heart block, time from last hospitalization, prematurity and interaction between heart block and prematurity were all observed to be statistically significant predictors of number of surgeries. For the children with more than two hospitalizations for cardiac related issues, the risk of an additional cardiac related surgery was six times greater than for children who had two or fewer hospitalizations for cardiac related issues. With respect to the statistically significant interaction term, when we conditioned on heart block specifically, premature children had approximately 1.5 times greater risk of having a subsequent surgery than children who were not born prematurely (RR = 1.48, 95% CI (0.84, 2.11)). The statistically significant time variable indicated that for each one unit increase in time since last hospitalization there was a 1.1% decrease in expected number of surgeries.

Binary portion of the model

In the binary (logistic) portion of the model, three variables emerged as statistically significant predictors of number of surgeries: time ($p = 0.01$), prematurity ($p = 0.05$), and heart block×prematurity interaction effect ($p = 0.05$). Although two of the variables sit squarely on the cusp of statistical

FIGURE 1
NUMBER OF SURGERIES FOR NON TRANSPLANT CHILDREN DIAGNOSED WITH ELECTROPHYSIOLOGICAL DISORDER

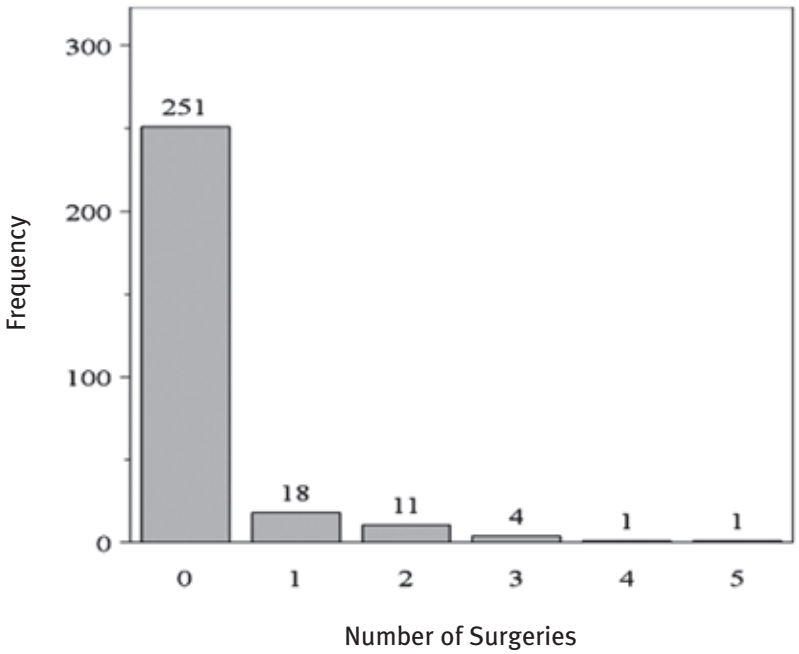
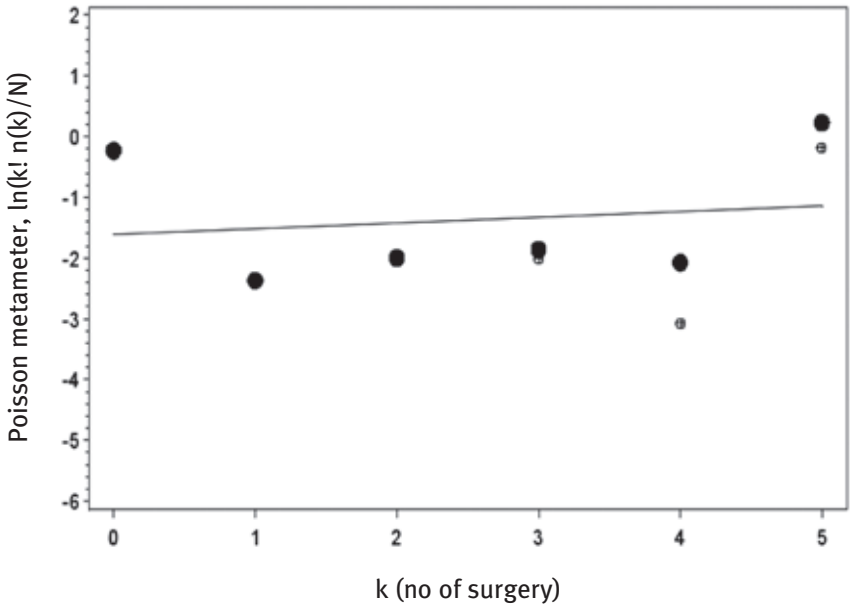


FIGURE 2
POISSONNESS PLOT OF THE CLINICAL DATA



truly significant predictors given the instructional nature of the example. It must be kept in mind that the interpretation of the binary portion of the model is different from the interpretation of the count portion. Although we are still trying to estimate the relationship between each of the clinical variables and a binary outcome, here the two levels of the binary variable consist of either structural (or true) zeroes or sampling zeroes that follow the Poisson distribution.

Consequently, the negative relationship between time since last hospitalization and the "no surgeries" portion of our outcome indicates an inverse relationship between time and "true" zeroes.

That is, as time increases (measured in months), there is a greater likelihood of a positive number of surgeries in the future. (OR = 0.92, 95% CI (0.87, 0.97)). The presence of a statistically significant interaction term indicates that premature children with heart block have a 60% higher odds of having surgery compared to non-premature children with heart block in this study population.

DISCUSSION

Appropriate modeling for number of surgeries in children diagnosed with EP disorders has some very obvious and important implications for clinicians and health service researchers. Most importantly, the presence of excess numbers of zeros in a dependent variable (i.e., number of surgeries) makes it extremely difficult to model using traditional parametric techniques.

This paper examines the utility of zero-inflated models, which are suitable when the excess number of zeros exceeds the number predicted by regular Poisson or negative binomial models. Hence, use of the more traditional Poisson or negative binomial models may actually lead to misleading inferences when interpreting the covariates of interest. Another clear advantage of zero-inflated modeling techniques is the ability to simultaneously examine the effects of covariates in both the zero and Poisson/negative binomial components of the model. When the zero-inflated Poisson model was applied to the clinical data, an interesting pattern emerged. A mixture model of two discrete distributions, logistic and Poisson, was deemed to be not only necessary, but crucial for analysis due to the complex nature of the distribution. Two of the most defining characteristics of this sample included an excessively high number of zeros (children without surgery) as well as a very skewed pattern to the discrete distribution (some children with as many as 2 through 5 surgeries). The characteristics of two very different distributions required the use of two different probability distributions for understanding the true nature of the sample data.

The process of choosing the best model is a trade-off between accuracy and simplicity. In this clinical example, while the Poisson modeling is the simplest model analytically, it underestimates the number of children who have not had any surgery. This is likely due to overdispersion, resulting from an excess number of zeros and a single Poisson parameter that is not sufficient to describe the population. A major assumption of the Poisson model is that the variance is equal to the mean. This assumption is violated in the presence of excess zeros in the data. To model the clinical data involving number of surgeries for children with EP disorders who have not had a heart transplant, the zero-inflated model was the clear choice based on the Vuong statistic discussed in the previous section. Both zero-inflated Poisson and zero-inflated negative binomial approaches resulted in many identical estimates, suggesting that once overdispersion has been accounted for, there were no other forms of heterogeneity in the sample data. Because zero-inflated Poisson was simpler than zero-inflated negative binomial model, we opted for the former.

There remain several notable limitations to this study. First, zero-inflated models involve complicated techniques not easily accessible to the clinicians. Second, they require estimation of a relatively large number of parameters. Third, there is no readily available software to assist the clinicians in analyzing this type of modeling technique, requiring outside assistance from a trained statistician.

Zero-inflated regression techniques were first introduced in the early 1990s' and continue to be used with increasing frequency in statistical methodology but have yet to gain a foothold in clinical sciences, especially in pediatric research. As zero-inflated modeling techniques become more widely used in medical research, clinicians will have more analytical tools at their disposal. Clinical science, by extension, will evolve beyond the existing set of general linear modeling techniques used today. The nature of mixture distributions will provide much more compelling and accurate results to all researchers, instead of those available through simple linear or log-linear modeling techniques.

An earlier version of this paper was presented at the 2010 Joint Statistical Meeting, Vancouver, BC.

References

[1] Lambert D. Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 1992; 34: 1-14

[2] Bohning D, Dietz E, Schlattmann P, et al. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society, Series A 1999; 162(2):195-209

[3] Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Statistics in Medicine 2002; 21:1461-9

[4] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine 2006; 25: 4279-92

[5] Hoaglin DC. A Poisonness plot. The American Statistician 1980; 34(3): 146-9

[6] Greene W. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. 1994; working paper, Department of Economics, New York University

[7] Greene W. Econometric Analysis, Prentice Hall. 2002

[8] Van Den Broek J. A score test for zero inflation in Poisson distribution. Biometrics 1995; 51: 738-43

[9] Vuong Q. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 1989; 57: 307-33

[10] Akaike H. A new look at the statistical model identification. IEEE Transactions on automatic control 1974; 19: 716-23

[11] Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge, UK: Cambridge University Press; 1998

[12] Long JS. Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications; 1997

[13] Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics 1986; 33: 341-65

APPENDIX A

The probability distribution of Poisson distribution is given by:

$$p(y_i \mid X = x) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}; \quad y_i = 0, 1, 2.....$$

$$with \quad \mu = e^{x'\beta}$$

where x denotes the covariates; and β denotes the vector of coefficients that need to be estimated.

The log-likelihood of the Poisson distribution is given by:

$$LL = \sum_{i=1}^{n}\left(y_i x'\beta - e^{x'\beta} - \ln y_i!\right)$$

The probability distribution of Negative Binomial distribution is given by:

$$\mathrm{P}\left(\mathrm{y}_i|\mu,\alpha\right) = \frac{\Gamma\left(\mathrm{y}_i + \frac{1}{\alpha}\right)}{\Gamma(\mathrm{y}_i+1)\Gamma\frac{1}{\alpha}}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\mu}{\frac{1}{\alpha}+\mu}\right)^{\mathrm{y}_i}$$

where α is the dispersion parameter.

The log-likelihood of the Negative Binomial distribution is given by:

$$LL = \sum\left\{\ln\Gamma(y_i+\frac{1}{\alpha}) - \ln\Gamma(y_i+1) - \ln\Gamma(\frac{1}{\alpha}) - \frac{1}{\alpha}\ln(1+\alpha\mu) + y_i\ln\left(\frac{\alpha\mu}{1+\alpha\mu}\right)\right\}$$

The probability density function of the Zero-inflated Poisson is given by:

$$\Pr\left(\mathrm{Y}_{\mathrm{i}} = 0\right) = \pi_i + (1-\pi_i)\,e^{-\mu_i}$$

$$\Pr\left(\mathrm{Y}_{\mathrm{i}} = y_i\right) = \left(1-\pi_i\right)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad y_i = 1, 2.....$$

where

$$\pi_i = \frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}} \quad ; \quad \mu_i = e^{x'\beta}$$

The log-likelihood of the Zero-inflated Poisson distribution is given by

$$LL = \sum_i\left[\begin{array}{l} I\left(y_i = 0\right)\ln\left(\frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}} + \frac{1}{1+e^{z_i'\gamma}}e^{-e^{x'\beta}}\right) + \\ I\left(y_i > 0\right)\ln\left(\frac{1}{1+e^{z_i'\gamma}} - e^{x'\beta} + y_i\left(x'\beta\right) - \Gamma\left(y+1\right)\right)\end{array}\right]$$

The probability density function of the Zero-inflated Negative Binomial is given by:

$$\Pr\left(Y_i = 0\right) = \pi_i + (1-\pi_i)\left(1+\alpha\mu_i\right)^{-\frac{1}{\alpha}}$$

$$\Pr\left(Y_i = y_i\right) = (1-\pi_i)\frac{\Gamma(y_i+\frac{1}{\alpha})}{y_i!\Gamma(\frac{1}{\alpha})}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\mu_i}{\frac{1}{\alpha}+\mu_i}\right)^{y_i} \qquad y_i = 1, 2.....$$

The log-likelihood of the Zero-inflated Negative Binomial is given by:

$$LL = \sum_i \left[\begin{array}{l} I\left(y_i = 0\right)\ln\left(\frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}} + \frac{1}{1+e^{z_i'\gamma}}\left(1+\alpha_i e^{x'\beta}\right)^{-\frac{1}{\alpha}}\right) + \\ I\left(y_i > 0\right)\ln\left(\frac{1}{1+e^{z_i'\gamma}} + \Gamma\left(y_i+\frac{1}{\alpha}\right) - \Gamma\left(y_i+1\right) - \Gamma\left(\frac{1}{\alpha}\right) + \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+e^{x'\beta}}\right)^{\frac{1}{\alpha}} + \left(\frac{e^{x'\beta}}{\frac{1}{\alpha}+e^{x'\beta}}\right)^{y_i}\right) \end{array} \right]$$

APPENDIX B

Poissonness plot

For a sample of N, the expected frequency from Poisson distribution can be written as:

$$\psi_y = N\frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0,1,2.....$$

For some fixed values of λ, if each observed frequency equals the expected frequency, then

$$\log\psi_y = \log(N) - \lambda + y\log\left(\lambda\right) - \log(y!)$$

If the data depicts poisson distribution, plotting log(Ψy) + log(y!) against y should yield a straight line with slope equals to log(λ) and intercept equals log(N) – λ.

APPENDIX C

The Lagrange Multiplier test statistic for overdispersion can be written as:

$$LM = \frac{\left(e'e - n\bar{y}\right)^2}{2\hat{\mu}'\hat{\mu}}$$

Where $e = y - \hat{y}$, and $\hat{\mu}$ is the predicted count of y under the Poisson model.

APPENDIX D

Van den Broek score test is based on the comparison of actual zeros to those predicted by the model. The test statistic can be written as :

$$Score = \frac{\left\{\sum_{i=1}^{n} I(y_i = 0) - \pi_{0i} / \pi_{0i}\right\}^2}{\sum_{i=1}^{n} \left(I(y_i = 0) - \pi_{0i}\right) / \pi_{0i} - n\bar{y}}$$

$I(y_i=0)$ is an indicator function with values equal to 1 if a given observation equals 0. Under the assumed poisson probability distribution, π_{0i} denotes the probability of observing zero for the i^{th} observation in the sample. The probability is allowed to vary by observation.

APPENDIX E

Sas code for Vuong test for Model fitting

```
proc nlmixed data=pcqli;

  parms   bll_0=0 bll_1=0 bll_2=0 bll_3=0 bll_4=0 bll_5=0   ;

  eta_lambda = bll_0 + bll_1*cbohoscr+ bll_2*hb +bll_3*prem
            + bll_4*timehosm +bll_5*hbprem ;

  lambda = exp(eta_lambda);
  loglike = hcs*log(lambda) -lambda - lgamma(hcs+1);

  model hcs ~ general(loglike);  predict _ll out=LL_1;

estimate "lambda " lambda;
predict lambda out = poi_out (rename = (pred = Yhat));

run;

proc nlmixed data=pcqli;

  parms bp_0=0 bp_1=0 bp_2=0 bp_3=0 bp_4=0 bp_5=0

        bll_0=0 bll_1=0 bll_2=0 bll_3=0 bll_4=0 bll_5=0;

  eta_prob = bp_0 + bp_1*cbohoscr+ bp_2*hb +bp_3*prem
           + bp_4*timehosm + bp_5*hbprem ;

  p_0 = exp(eta_prob)/(1 + exp(eta_prob));

  eta_lambda =bll_0 + bll_1*cbohoscr + bll_2*hb +bll_3*prem +
            bll_4*timehosm +bll_5*hbprem ;
```

```
lambda = exp(eta_lambda);

  if hcs=0 then prob = p_0 + (1-p_0)*exp(-lambda);
  if hcs=0 then loglike = log(prob);
  else loglike = log(1-p_0) + hcs*log(lambda) -lambda - lgamma(hcs+1);
  model hcs ~ general(loglike);

predict eta_lambda out = zip_out1 (keep = pred ahci rename = (pred = Yhat));
predict p_0 out = zip_out2 (keep = pred rename = (pred = p_0));
predict _ll out=LL_2;

run;

/*  Vuong Test : ZIP VS. Poisson *********/

title1 'Vuong test for ZIP vs. poisson';
title2 'H0 = no improvement of ZIP over poisson';
data ll_diff;
  merge ll_1 (rename= (pred=ll_poisson))
                ll_2 (rename= (pred=ll_zip));
run;
data ll_diff;
  set ll_diff;
  lr_i = ll_zip - ll_poisson;
  keep ll_poisson ll_zip lr_i;
run;
proc means data=ll_diff vardef=n;
  var lr_i;
  output out=vuong_stats mean=LR var=V_lr_i n=n;
run;
data vuong_stats;
  set vuong_stats;
  Vuong = (LR /sqrt(V_lr_i/n));
  p = 2*(1-probnorm(vuong));
  put vuong= p=;
run;
```

*