Analysis of count data in the setting of cervical cancer detection

Christina G Bracamontes,[1] Thelma Carrillo,[1] Jane Montealegre,[2] Leonid Fradkin,[3] Michele Follen,[4] Zuber D Mulla [1,5]

[1]Department of Obstetrics and Gynecology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center El Paso, El Paso, Texas, USA
[2]Department of Pediatrics, and Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA
[3]Department of Obstetrics and Gynecology, Brookdale University Hospital and Medical Center, Brooklyn, New York, USA
[4]Department of Obstetrics, Gynecology, and Women's Health, Kings County Hospital, Brooklyn, New York, USA
[5]Office of Faculty Development, Texas Tech University Health Sciences Center El Paso Paul L Foster School of Medicine, El Paso, Texas, USA

Correspondence to
Dr Zuber D Mulla, Office of Faculty Development, Texas Tech University Health Sciences Center El Paso Paul L Foster School of Medicine, El Paso, Texas, USA; zuber.mulla@ttuhsc.edu

The abstract was presented as a poster (number LB05) at the Annual Meeting of the American College of Epidemiology, in Decatur, Georgia, on September 28, 2015.

Accepted 23 June 2020
Published Online First 13 July 2020

To cite: Bracamontes CG, Carrillo T, Montealegre J, *et al*. *J Investig Med* 2020;**68**:1196–1198.

ABSTRACT

Women with an abnormal Pap smear are often referred to colposcopy, a procedure during which endocervical curettage (ECC) may be performed. ECC is a scraping of the endocervical canal lining. Our goal was to compare the performance of a naïve Poisson (NP) regression model with that of a zero-inflated Poisson (ZIP) model when identifying predictors of the number of distress/pain vocalizations made by women undergoing ECC. Data on women seen in the colposcopy clinic at a medical school in El Paso, Texas, were analyzed. The outcome was the number of pain vocalizations made by the patient during ECC. Six dichotomous predictors were evaluated. Initially, NP regression was used to model the data. A high proportion of patients did not make any vocalizations, and hence a ZIP model was also fit and relative rates (RRs) and 95% CIs were calculated. AIC was used to identify the best model (NP or ZIP). Of the 210 women, 154 (73.3%) had a value of 0 for the number of ECC vocalizations. NP identified three statistically significant predictors (language preference of the subject, sexual abuse history and length of the colposcopy), while ZIP identified one: history of sexual abuse (yes vs no; adjusted RR=2.70, 95% CI 1.47 to 4.97). ZIP was preferred over NP. ZIP performed better than NP regression. Clinicians and epidemiologists should consider using the ZIP model (or the zero-inflated negative binomial model) for zero-inflated count data.

INTRODUCTION

Count data arise frequently in the health sciences. For example, researchers accessing various state health databases may encounter variables that are counts. A tumor registry may record the number of regional nodes that are positive for a malignancy, while a birth defect registry may capture the number of previous live births or the number of prenatal care visits. The Poisson regression model is typically used by epidemiologists when the outcome is a count variable.[1 2] While data analysts may be tempted to use linear regression to analyze count data, Poisson regression has the advantage of being perfectly suited to dealing with a dependent variable that is discrete and frequently has a highly skewed distribution.[2]

The mean and variance of the Poisson distribution are equal[3]; however, for many datasets, the observed variance is greater than the assumed variance, a phenomenon known as overdispersion.[2 4] A departure from the Poisson model can occur if the observed data consist of an excessive number of zeros.[4] Overdispersion in the setting of excessive zero counts will result in an underestimate of the variance of the parameter estimate.[4]

The zero-inflated Poisson (ZIP) probability distribution may be used to adjust for overdispersion when the occurrence of zeros is greater than expected.[4] Students of epidemiology are typically introduced to the naïve Poisson (NP) regression model during an intermediate course but may not encounter the ZIP model since popular intermediate epidemiology textbooks do not address the topic of analyzing zero-inflated count data.[1 5 6]

The objective of our investigation was to compare the performance of the NP regression model with the ZIP regression model when analyzing count data from our study of the development and application of a novel multispectral digital colposcope.[7] A colposcope is an instrument used by a healthcare provider to closely examine the uterine cervix and other portions of the female reproductive tract for signs of disease (colposcopy). Women who have an abnormal Pap smear (a screening test for cervical cancer) are typically referred to colposcopy. During colposcopy, the provider may perform an endocervical curettage (ECC), which is a scraping of the lining of the endocervical canal. A typical ECC is completed in a matter of seconds and may cause varying degrees of discomfort or pain. The outcome variable of interest for our analysis was the number of pain or distress vocalizations made by the patient during the performance of the ECC.

MATERIALS AND METHODS

Source of subjects

Subjects were recruited from patients who were seen in the colposcopy clinic of the Department of Obstetrics and Gynecology, Texas Tech University Health Sciences Center El Paso (El Paso, Texas) between December 2012 and December 2014. The number of ECC pain vocalizations was collected for subjects

who were enrolled in the study between May 2013 and September 2014. Subjects were also administered a survey in the language of their choice, English or Spanish.

ZIP model

We assumed that our population consisted of two types of subjects. The first type gave rise to counts that follow the Poisson distribution, which may contain zeros.[8] The second type always gives a zero count. The ZIP model has two parts: first, a logit model is created for the subjects who always produce zeros (the zero probability model), and second, a Poisson model is generated predicting the counts for those subjects who do not always produce zeros.[4 9] The reader is referred elsewhere for the mean and variance of a random variable that has a ZIP distribution.[4 8]

Data analysis

Data were analyzed using SAS V.9.3. The outcome was the number of pain/distress vocalizations made during ECC. The number of pain/distress vocalizations was recorded by a member of the research team who was present during the colposcopy. The process of creating and refining the data collection tool that was used to record self-reported pain, self-reported anxiety, and the number of vocalizations heard by the research team member took several months, and hence some subjects have missing values for the number of distress vocalizations. This pain/anxiety data collection tool was created to evaluate any distress that may have been caused by the novel multispectral digital colposcope and was modified during the course of the trial in order to measure pain at every step of the colposcopy.

Predictors for inclusion in the regression models were chosen based on clinical or epidemiological relevance. Six dichotomous predictor variables were created: age (≥30 years vs <30 years), language in which the subject took the survey (English vs Spanish), history of sexual abuse (yes vs no), number of vaginal deliveries (≥1 vs 0), smoked ≥100 cigarettes during lifetime (yes vs no), and length of the colposcopy (>13 min vs ≤13 min). Thirteen minutes was the median value for the length of the colposcopy in our sample.

NP and ZIP regression models are examples of generalized linear models. The data analyst must specify both a link function and a probability distribution in order to fit a generalized linear model. The GENMOD procedure was used to fit a single NP regression model and a single ZIP regression model with a log link specified for both models. (The default link function when either a Poisson or ZIP probability distribution is specified is log base *e*.) The six independent variables noted earlier were entered simultaneously, and no variable selection procedure was used. For the ZIP model, the zero-inflation probability, that is, the probability of zero counts in excess of the frequency predicted by the underlying distribution, was requested using the PZERO keyword.[4 8]

To account for varying time spans (eg, lengths of follow-up) when fitting Poisson regression models, an offset variable is required.[2] For most patients, an ECC lasts 10 s. Since the length of an ECC does not vary substantially between patients, an offset variable (ln or length of the ECC) was not required for our analyses. The relative rate (RR) of ECC pain vocalization for each of the predictor variables was calculated by taking the antinatural logarithm of the regression parameter estimate. Adjusted RRs were reported along with 95% CIs and p values.

A large ratio of the deviance to the df was used to identify the possibility of the presence of overdispersion when fitting our NP model.[2] Akaike's information criterion (AIC) is frequently used to compare the relative fit of two or more models with a lower value corresponding to the more desirable model.[2] The AIC measures how well the model fits the data while adjusting for the number of variables in the model.[10] When comparing the AIC from two or more regression models, a larger value of the AIC may indicate that the model in question has poorer out-of-sample predictive ability.[10] In our study, the AIC was used to determine which was the preferred model, the NP or the ZIP. The AICs from both models were compared. The model with the smaller value was considered to be the better fitting model.

Table 1 Characteristics of 210 women who underwent ECC during colposcopy, El Paso, Texas

Characteristics	Number (%)
Predictor variables	
Age ≥30 years	120 (57.1)
Language in which survey was administered	
English	126 (60.0)
Spanish	84 (40.0)
History of sexual abuse	
Yes	8 (3.8)
No	202 (96.2)
Had ≥1 vaginal deliveries	147 (70.0)
Smoked ≥100 cigarettes during lifetime	58 (27.6)
Colposcopy lasted >13 min*	99 (47.1%)
Outcome	
Count of distress/pain vocalizations during ECC	
0	154 (73.3)
1	23 (11.0)
2	20 (9.5)
3	6 (2.9)
4	3 (1.4)
5	3 (1.4)
6	1 (0.5)

*Length of colposcopy ranged from 3 to 34 min (mean=14.7 min, median=13.0 min).
ECC, endocervical curettage.

RESULTS

A total of 470 women were enrolled in the multispectral digital colposcopy trial. Data on pain, anxiety, and distress vocalizations were available for 283 subjects. Of the 283, 247 underwent ECC. The records of 210 women had complete data for the variables included in the current analysis.

Selected characteristics of the sample are reported in table 1. Sixty percent of the women completed the accompanying survey in English. A large proportion (73.3%) had a value of 0 for the outcome (the number of pain vocalizations made during the ECC).

Table 2 Adjusted* RRs, SEs, 95% CIs, and p values for distress/pain vocalizations from a NP regression model and a ZIP regression model (N=210)

	NP				ZIP			
Variable	**RR**	**SE**	**95% CI**	**P value**	**RR**	**SE**	**95% CI**	**P value**
Age ≥30 years vs <30 years	1.35	0.2303	0.86 to 2.11	0.20	1.35	0.2815	0.78 to 2.34	0.29
English survey versus Spanish survey	0.49	0.2080	0.32 to 0.73	0.001	0.62	0.2498	0.38 to 1.01	0.06
Sexual abuse (yes vs no)	3.26	0.2962	1.82 to 5.82	<0.0001	2.70	0.3107	1.47 to 4.97	0.001
Had ≥1 vaginal deliveries vs 0	1.05	0.2461	0.65 to 1.70	0.85	1.07	0.2973	0.60 to 1.92	0.81
Smoked ≥100 cigarettes during lifetime (yes vs no)	1.23	0.2021	0.83 to 1.83	0.30	1.13	0.2341	0.71 to 1.78	0.61
Colposcopy lasted >13 min (vs ≤13 min)	2.02	0.1950	1.38 to 2.97	0.0003	1.48	0.2339	0.94 to 2.34	0.09

*Each RR is adjusted for the remaining variables found in the model.
NP, naïve Poisson; RR, relative rate; ZIP, zero-inflated Poisson.

Adjusted RRs for pain vocalizations made during ECC are shown in table 2. The ratio of the deviance to the df was 1.44 for the NP model. The NP model identified three statistically significant variables: language in which the survey was administered (English compared with Spanish, RR=0.49, 95% CI 0.32 to 0.73), history of sexual abuse (RR=3.26, 95% CI 1.82 to 5.82), and the length of the colposcopy (>13 min compared with ≤13 min, RR=2.02, 95% CI 1.38 to 2.97). The AIC from the NP model was 446.1. The ZIP model identified one statistically significant variable: history of sexual abuse (women who had been abused had 2.7 times the rate of vocalization than non-abused women, 95% CI 1.47 to 4.97). The AIC from the ZIP model was 400.2, indicating that the ZIP was the preferred model.

Using the PZERO keyword in SAS, the data analyst can request that the zero-inflation probability be reported in the output. The zero-inflation probability is the probability of zero counts in excess of the frequency predicted by the Poisson distribution.[8] In our study, the zero-inflation probability was 0.595 or 59.5%.

DISCUSSION

Our analysis revealed that women who had a history of being sexually abused were more likely than women who had not suffered sexual abuse to verbally express pain during the ECC. We could not find any published investigations of the possible association between a history of sexual abuse and pain/distress in patients undergoing colposcopy. We also found that ZIP performed better than NP regression when modeling zero-inflated count data from our cervical cancer detection study. In the presence of a high proportion of zero counts, the phenomenon of overdispersion can result in spuriously low SEs and p values when NP is used,[4] and this is the likely explanation for why our NP model identified three RRs as being statistically significant, while the ZIP model identified only one.

Overdispersion is a common occurrence when fitting Poisson regression models. When analyzing count data and, in particular, zero-inflated count data, clinicians and epidemiologists should be aware of alternatives to the NP model such as the ZIP. Other options include using the negative binomial and the zero-inflated negative binomial distributions, and the reader is referred elsewhere for information on these models.[2 4 11]

Acknowledgements The authors thank the clinic staff in the Department of Obstetrics and Gynecology, Texas Tech University Health Sciences Center El Paso, for their assistance with data collection and data entry.

Contributors MF offered overall project leadership. CGB, TC, LF, and MF collected the data. TC and CGB managed the data. ZDM analyzed the data. ZDM drafted the initial manuscript. All authors offered input on the study results and offered critical input and review, which led to the creation of the final version of the manuscript.

Funding This study was supported in part by the National Institutes of Health award P01 CA082710-13 (Optical Technologies and Molecular Imaging for Cervical Neoplasia) and the Department of Obstetrics and Gynecology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center El Paso, El Paso, Texas.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Our study was approved by the Institutional Review Board for the Protection of Human Subjects, Texas Tech University Health Sciences Center at El Paso (approval number E12117).

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
Zuber D Mulla http://orcid.org/0000-0003-1670-5702

REFERENCES

1 Szklo M, Nieto FJ. *Epidemiology beyond the basics*. Gaithersburg, Maryland: Aspen Publishers, Inc, 2000.
2 Allison PD. *Logistic regression using the SAS® system: theory and application*. Cary, North Carolina: SAS Institute, Inc, 1999.
3 Daniel WW. *Biostatistics: a foundation for analysis in the health sciences*. 5th ed. New York: John Wiley & Sons, Inc, 1991..
4 Lee J-H, Han G, Fulp WJ, *et al*. Analysis of overdispersed count data: application to the human papillomavirus infection in men (HIM) study. *Epidemiol Infect* 2012;140:1087–94.
5 Koepsell TD, Weiss NS. *Epidemiologic methods: studying the occurrence of illness*. New York: Oxford University Press, Inc, 2003.
6 Kelsey JL, Whittemore AS, Evans AS, *et al*. *Methods in observational epidemiology*. 2nd ed. New York: Oxford University Press, Inc, 1996..
7 Buys TPH, Cantor SB, Guillaud M, *et al*. Optical technologies and molecular imaging for cervical neoplasia: a program project update. *Gend Med* 2012;9:S7–24.
8 SAS Institute. Zero-Inflated Poisson models. SAS/STAT® 9.2 user's guide, second edition. Available: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect042.htm [Accessed 18 May 2020].
9 Institute for Digital Research & Education, UCLA. Zero-Inflated poisson regression R data analysis examples. Available: https://stats.idre.ucla.edu/r/dae/zip/ [Accessed 19 May 2020].
10 Rothman KJ, Greenland S, Lash TL. *Modern epidemiology third edition*. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.
11 Armitage P, Colton T. *Encyclopedia of biostatistics*. Second Edition. Chichester, England: John Wiley & Sons, Ltd, 2005.