Journal of Scientific Research & Reports
3(17): 2286-2302, 2014; Article no. JSRR.2014.17.004

SCIENCEDOMAIN *international*
www.sciencedomain.org

Models for Injury Count Data in the U.S. National Health Interview Survey

Jin Peng[1,2], Tianmeng Lyu[2,3], Junxin Shi[2], Haikady N. Nagaraja[1] and Huiyun Xiang[1,2*]

[1]The Ohio State University, College of Public Health, Columbus, Ohio, USA.
[2]Center for Injury Research and Policy, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio, USA.
[3]Peking University, School of Mathematical Sciences, Department of Probability and Statistics, Beijing, P.R. China.

Authors' contributions

This work was carried out in collaboration between all authors. Author JP managed the literature search, performed the statistical analysis, wrote the first draft of the manuscript, edited and sent it to the journal for publication. Author JS developed the preliminary SAS codes. Author TL contributed to data analysis and results interpretation. Author HNN supervised the study, reviewed results and participated in drafting the manuscript. Author HX conceived and designed the study. He is the author for correspondence. All authors read and approved the final manuscript submitted for publication.

Original Research Article

Received 12th February 2014
Accepted 31st May 2014
Published 17th July 2014

ABSTRACT

Aims: To examine the best count data model for injury data in the National Health Interview Survey (NHIS). To compare the best count data model with traditional logistic regression model in analyzing injury data in NHIS.
Data Source: 2006-2010 medically consulted non-occupational injury data from National Health Interview Survey (NHIS).
Methodology: Six count data models (Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated NB (ZINB), hurdle Poisson (HP), and hurdle NB (HNB)) were compared using Likelihood Ratio (LR) test and Vuong test. Injury count was used as the dependent variable in count data models. Independent variables included age, gender, marital status, race, education, poverty status, disability status and medical insurance coverage status. Dichotomized injury count was used as the dependent variable in logistic

**Corresponding author: E-mail: huiyun.xiang@nationwidechildrens.org;*

regression model. The same independent variables used in count data models were included in logistic regression model. The model fit of logistic regression was examined by Hosmer and Lemeshow goodness of fit test.

Results: Among 248,850 participants aged 18-64, 98.37% have no medically consulted non-occupational injuries, 1.55% have 1 medically consulted non-occupational injury, 0.07% have 2 or more medically consulted non-occupational injuries. Zero-inflated negative binomial (ZINB) model offered the best fit. Logistic regression model provided a good fit but resulted in different estimates from ZINB model.

Conclusion: Zero-inflated negative binomial (ZINB) model demonstrated the potential to be the best model for injury count data with excess zeros. Given the infrequent occurrence of multiple injuries in our data, the logistic regression model is appropriate for assessing injury burden and identifying injury risks. However, for more frequently-occurring injuries (e.g. sports injuries), logistic regression may undercount the total number of injuries and result in biased estimates. The evaluation procedure and model selection criteria presented in this paper provide a useful approach to modeling injury count data with excess zeros.

Keywords: Injury epidemiology; National Health Interview Survey (NHIS); injury count data; logistic regression model; Zero-inflated Negative Binomial (ZINB) model.

1. INTRODUCTION

Count outcomes are commonly encountered in injury epidemiology, such as the count of occupational injuries among US workers and the count of falls among elderly people. Most injury studies [1-4] use logistic regression models to analyze injury count data. In logistic regression, injury count is dichotomized into a binary outcome (absence or presence of injuries). This approach may underestimate the burden of injuries, diminish the accuracy of identifying injury risk factors, and bias evaluations of injury intervention programs. Models for analyzing count data have been developed [5]. Because count data models use exact injury count as the dependent variable, they are capable of assessing injury burden in terms of injury frequency rather than the presence or absence of injuries.

Even though count data models are perhaps better suited to handle injury count data than logistic regression, few applications can be found in the current literature. Karazsia and van Dulmen [6] examined the appropriateness of four count data models (Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB)) in identifying predictors of children's medically attended injuries. They found that some models fit the data more accurately than others and the predictors of children's medically attended injuries tended to vary depending on the specific model utilized. Therefore, they encouraged researchers to select count data models according to the characteristics of their data. Ullah et al. [7] compared four count data models (Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB)) in analyzing falls count data from four separate datasets. Their results showed that the NB model offered the best fit and therefore was recommended for future studies in modeling falls count data. Khan et al. [8] examined the appropriateness of six count data models (Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), hurdle Poisson (HP) and hurdle negative binomial (HNB)) in analyzing falls count data from a prospective cohort study. They found that NB-based regression models performed better than other count data models, with the HNB model offering the best fit. They did however point out that the HNB model might not offer the same advantage for other falls count data.

National Health Interview Survey (NHIS) has been widely used in injury research [9-13]. Although injury count data are available in this survey, no study has examined the best count data model for these data. This study aimed to identify the best count data model for the injury data in NHIS and compare the best count data model results with traditional logistic regression model results. For these purposes, six count data models (Poisson, negative binomial (NB), hurdle Poisson (HP), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB)) were implemented and compared. The best count data model was then compared to logistic regression model in terms of model fit and regression estimates.

2. MATERIALS AND METHODS

2.1 Data Source

National Health Interview Survey (NHIS) has been widely used in injury epidemiology. It is publicly available on the website of National Center for Health Statistics (NCHS) http://www.cdc.gov/nchs/nhis.htm [14], which is part of the Centers for Disease Control and Prevention (CDC) in the United States. With a complex, multistage sampling design, NHIS is a cross-sectional survey conducted by the National Center for Health Statistics through personal household interviews. Health information is collected on all members of selected households who are at home at the time of the interview; for children, adults who are not at home and those are physically or mentally unable to respond, information is provided by a knowledgeable adult family member. The overall response rate for the survey is close to 90 percent. The last major revisions of the content and the sampling design occurred in 1997 and 2006, respectively. Since there were no major changes in the content or in the sampling design from 2006 to 2010, we could develop models using the combined 5-year data (2006-2010). The NHIS data are segmented into several files based on which respondent was interviewed. In this study, we used the person file (includes persons of all ages) for demographics, including disability information; the family file for family income; the adult file (includes only persons aged 18 years or older) for occupation; and the injury episode file for injury characteristics.

The data analyzed in this study are de-identified publicly accessible data. The Institutional Review Board of Nationwide Children's Hospital reviewed the study protocol and decided that this study was exempted.

2.2 Data Analysis

Poisson, negative binomial (NB), hurdle Poisson (HP), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB) models were each fit to the data using PROC NLMIXED in SAS 9.3 (SAS Institute, Cary, NC). The dependent variable was the count of medically consulted non-occupational injuries that occurred in the last three months (prior to the interview) among people between 18 and 64 years old. To obtain graphical illustration of model fit, the observed proportions minus mean predicted probabilities of each injury count were plotted (using intercept-only models). Independent variables included in the models were age, sex, marital status, race, education level, poverty status, disability status and medical insurance coverage status. For the purposes of this study, continuous variable age was categorized into three age groups. Ages 18-29 years are young adults, ages 30-54 years are adults, and ages 55-64 years are older adults. Other independent variables were categorical variables as defined in the survey questionnaire. The

same set of independent variables was used in the logit part (See Appendix 1 function (6), (9)) and the log linear part (See Appendix 1 function (2)) of HP, HNB, ZIP and ZINB models.

Various statistical tests were conducted to examine over-dispersion and to compare model fit. Over-dispersion in the Poisson regression was evaluated by Cameron and Trivedi test [15]. Since NB model and HP model can reduce to the Poisson model under certain conditions, they are nested models and can be compared using the Likelihood Ratio (LR) test. For those non-nested models, such as Poisson and ZIP, Vuong test [16,17] was used to compare their fit to our data.

Coefficients of the best count data model were estimated in STATA 12 (StataCorp, College Station, TX) with the incorporation of survey weights. In order to compare the best count data model with logistic regression model, a logistic regression (with the same set of independent variables as in the best count data model) was implemented. Dichotomized injury count ("0" indicating no injuries and "1" indicating one or more injuries) was used as the dependent variable in logistic regression model. The model fit of logistic regression to our data was examined by Hosmer and Lemeshow goodness of fit test [18].

3. RESULTS

3.1 Descriptive and Graphical Analysis of Injury Count Data in NHIS

Table 1 presents the distribution of the dependent and independent variables used in our models. Among 248,850 participants, 98.37% reported no medically consulted non-occupational injuries, 1.55% reported 1 injury and 0.07% reported 2 or more injuries. The hurdle negative binomial (HNB) model could not be fitted to our data. We tried to fit the data with HNB model using both SAS 9.3 (SAS Institute, Cary, NC) and STATA 12 (StataCorp, College Station, TX). But the model never converged. To verify if our SAS programs on HNB model were accurate, we applied the SAS programs from another similar study [19] to our data but the model still did not converge. We were able to implement and compare the other five count data models in our study.

Fig. 1 shows the observed proportions minus the mean predicted probabilities of each injury count for five count data models. The Poisson model predicted many fewer 0s and many more 1s than observed. The NB, HP, ZIP and ZINB models showed a substantial improvement over the Poisson model. Although these models each predicted similar proportions of 0s as observed, they tended to underestimate the proportion of 1s and overestimate the proportion of 2s.

3.2 Count Data Model Selection

The observed variance of non-occupational injury counts (0.020) was larger than the mean (0.017), indicating evidence of over-dispersion. This is supported by the results of Cameron and Trivedi test ($t = 9.12$, $P<.0001$). Based on the results from Likelihood Ratio (LR) tests and Vuong tests, zero-inflated negative binomial (ZINB) model resulted in the best statistical fit (see Table 2). Although it was difficult to determine which model fits better among negative binomial (NB), hurdle Poisson (HP) and zero-inflated Poisson (ZIP) models (given $|V|<1.96$, none of the models were preferred), they performed better than the Poisson model.

Table 1. Distribution of the dependent and independent variables in count data models

Variable name	Frequency	Percent
Dependent variable		
Number of non-occupational injuries		
0	244792	98.37
1	3855	1.55
2	160	0.06
3	27	0.01
4	10	0
5	6	0
Independent variables		
Age group		
18-29	65022	26.13
30-54	141032	56.67
55-64	42796	17.20
Gender		
Male	119620	48.07
Female	129230	51.93
Marital Status		
Married	138623	55.71
Single/never married	72098	28.97
Separated/divorced/widowed	36148	14.53
Unknown	1981	0.80
Race		
Hispanic	58795	23.63
Non-Hispanic White	132908	53.41
Non-Hispanic Black	37548	15.09
Others	19599	7.88
Education		
No college education	108836	43.74
Received College education[1]	133439	53.62
Unknown	6575	2.64
Poverty status		
Poor	29402	11.82
Near poor	36535	14.68
Not poor	140851	56.60
Unknown	42062	16.90
Disability status		
Non-disabled	223611	89.86
Disabled[2]	24284	9.76
Unknown	955	0.38
Medical insurance coverage status		
Not covered with medical insurance	57632	23.16
Covered with medical insurance	188465	75.73
Unknown	2753	1.11

1. Defined as "receipt of some college education" or "receipt of bachelor's degree or above".

2. Defined as "limited, caused by at least one chronic condition", "Limited, not caused by chronic condition" or "Limited, unknown if condition is chronic"

Table 2. Model comparison results of the fitted models for injury count data from 2006-2010 NHIS

Nested models	**LR statistics[a]**	**Preferred model**
Poisson vs NB	507	NB
Poisson vs HP	489	HP
Non-nested models	**Vuong statistics[b]**	**Preferred model**
HP vs NB	-1.04	N/A
ZIP vs NB	-1.01	N/A
ZIP vs HP	61.2	ZIP
ZIP vs Poisson	7.24	ZIP
ZINB vs NB	4.23	ZINB
ZINB vs HP	61.3	ZINB
ZINB vs ZIP	3.69	ZINB
ZINB vs Poisson	7.54	ZINB

NB, negative binomial; HP, hurdle Poisson; ZIP, zero-inflated Poisson; ZINB, zero-inflated negative binomial; NHIS, National Health Interview Survey.

a. Likelihood Ratio (LR) test (for nested models): If the LR statistics $\chi^2 > \chi_1^2 = 3.84$, the more complex model is preferred.

b. Vuong (V) test (for non-nested models): If $V > 1.96$, the first model is preferred; if $V < -1.96$, then the second model is preferred; if $|V| < 1.96$, none of the models are preferred.

3.3 Estimation Results from the Best Count Data Model

Results from the best count data model (ZINB model) are shown in Table 3. The first column describes the independent variables used in the ZINB model. The second and third columns report the estimated coefficients and p-values in the logit part of the ZINB model (see Appendix 1 function (9)). The last two columns report the estimated coefficients and p-values in the log linear part of the ZINB model (see Appendix 1 function (2)). Based on these estimated coefficients, the ZINB model can be used to predict the probability of a person being in the always-zero group and his/her expected number of injuries given that he/she is in the not-always-zero group. For example, consider a 30 year old White married man who is college educated, not poor and has medical insurance. If he has disabilities, the probability of him being in the always-zero group is 50% and the expected number of injuries given he is in the not-always-zero group is 0.054. If he does not have disabilities, the probability of him being in the always-zero group remains the same (50%), whereas the expected number of injuries given he is in the not-always-zero group would be 0.020. These results have two important implications. First, disability status may not be associated with the probability of a person being in the always-zero group. Second, given in the not-always-zero group, people with disabilities would sustain approximately three times as many injuries as their non-disabled counterparts. Furthermore, the ZINB model can also be used to estimate the odds ratio of getting injured between people with and without disabilities. Another ZINB model (a single covariate disability was included in the log linear part; age, marital status, race, education level, poverty status and medical insurance coverage status were included in the logit part) was implemented and compared with another logistic regression model (a single covariate disability was included in the model). The point estimate of the odds ratio obtained from the ZINB model (OR=2.91) was approximately the same as from the logistic regression model (OR=3.01) (See Appendix 2 for the formulas of calculating odds ratios). The statistical theory of building confidence intervals for odds ratios resulting from ZINB model is not available in the current literature.

Table 3. Estimated coefficients from the ZINB regression for injury count data from 2006-2010 NHIS

Variable	ZINB model coefficient estimates			
	Logit part[1]		Log linear part[2]	
Constant	0.71	(*P*= .57)	-4.26*	(*P*= .00)
Ages 18-29 years (reference group)				
Ages 30-54 years	0.95	(*P*= .27)	-0.17	(*P*= .19)
Ages 55-64 years	-0.38	(*P*= .88)	-0.19	(*P*= .15)
Female (reference group)				
Male	-0.16	(*P*= .80)	0.02	(*P*= .71)
Married (reference group)				
Single/never married	-0.15	(*P*= .88)	0.25*	(*P*= .03)
Separated/divorced/widowed	-1.33	(*P*= .08)	0.37*	(*P*= .00)
Hispanic (reference group)				
Non-Hispanic White	-2.00	(*P*= .16)	0.33*	(*P*= .04)
Non-Hispanic Black	-2.28*	(*P*= .04)	-0.002	(*P*= .99)
Other races	-0.31	(*P*= .76)	0.40	(*P*= .12)
Not received college education (reference group)				
Received college education	-1.01	(*P*= .12)	0.17*	(*P*= .00)
Poor (reference group)				
Near poor	-0.22	(*P*= .71)	-0.25*	(*P*= .03)
Not poor	-2.47	(*P*= .07)	-0.47*	(*P*= .01)
Unknown poverty status	-0.55	(*P*= .50)	-0.63*	(*P*= .00)
Not disabled (reference group)				
Disabled	-1.23	(*P*= .35)	1.01*	(*P*= .00)
Not covered with medical insurance (reference group)				
Covered with medical insurance	0.12	(*P*= .75)	0.30*	(*P*= .00)

**indicates significant estimated coefficients (P < .05)*

ZINB, zero-inflated negative binomial; NHIS, National Health Interview Survey

$$1.\ Log\left(\frac{p_i}{1-p_i}\right) = Z_i'\gamma$$

$$2.\ Log\ (\mu_i) = X_i'\beta$$

3.4 Estimation Results from the Logistic Regression Model

Table 4 provides the results from a logistic regression model with the same independent variables as in the ZINB model. It shows that receiving a college education, having disabilities and being covered with medical insurance are all associated with an increased odds of reporting one or more medically consulted non-occupational injuries. Younger people (age 18-29) were more likely to report one or more medically consulted non-occupational injuries than older people (age 30-64); married people were less likely to report one or more medically consulted non-occupational injuries than unmarried people; Hispanics were less likely to report one or more medically consulted non-occupational injuries than other races; poor people were more likely to report one or more medically consulted non-occupational injuries than near poor or not poor people. The results from Hosmer and Lemeshow goodness of fit test indicated that this logistic regression model resulted in a good fit to our data (*P* = .30). This is likely attributable to the rare occurrence of multiple injuries in our data (only 0.07% of the participants had two or more injuries, see Table 1).

Table 4. Estimated coefficients from the logistic regression for injury count data from 2006-2010 NHIS

Variable	Logistic regression coefficient estimates	
Constant	-4.84*	(*P*= .00)
Ages 18-29 years (reference group)		
Ages 30-54 years	-0.23*	(*P*= .00)
Ages 55-64 years	-0.19*	(*P*= .00)
Female (reference group)		
Male	0.02	(*P*= .50)
Married (reference group)		
Single/never married	0.28*	(*P*= .00)
Separated/divorced/widowed	0.44*	(*P*= .00)
Hispanic (reference group)		
Non-Hispanic White	0.65*	(*P*= .00)
Non-Hispanic Black	0.34*	(*P*= .00)
Other races	0.46*	(*P*= .02)
Not received college education (reference group)		
Received college education	0.26*	(*P*= .00)
Poor (reference group)		
Near poor	-0.23*	(*P*= .00)
Not poor	-0.25*	(*P*= .00)
Unknown poverty status	-0.56*	(*P*= .00)
Not disabled (reference group)		
Disabled	1.01*	(*P*= .00)
Not covered with medical insurance (reference group)		
Covered with medical insurance	0.29*	(*P*= .00)

**indicates significant estimated coefficients (P<.05). NHIS, National Health Interview Survey*

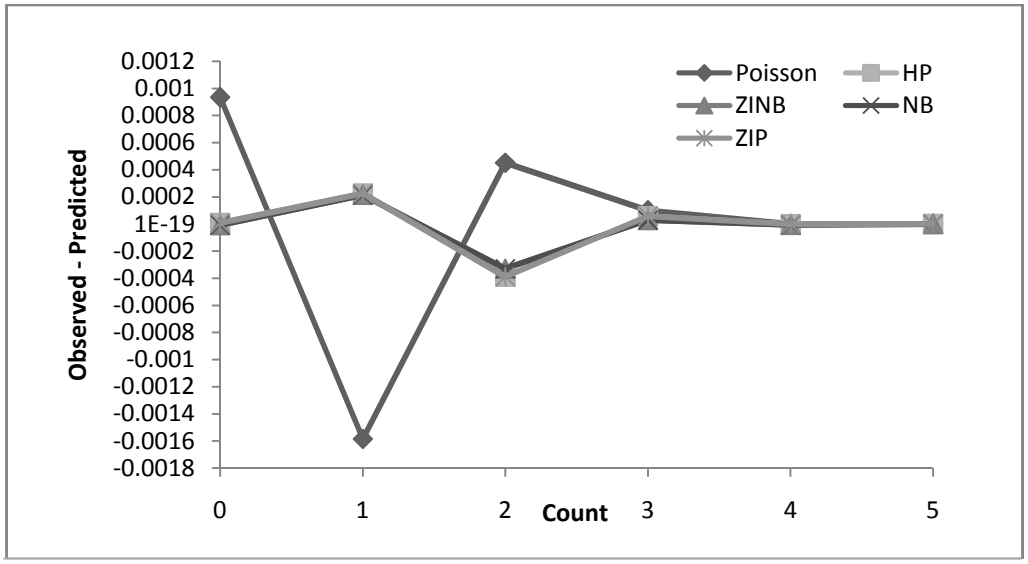

Fig. 1. Observed minus predicted probabilities at each injury count: Poisson, negative binomial (NB), hurdle Poisson (HP), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB) models; while the non-Poisson models produced close results, Vuong test showed clear superiority of the ZINB model over others

4. DISCUSSION

Injuries are a major public health concern in the United States. Despite ongoing progress in injury prevention, the toll of injuries in terms of medical expenses and work loss remains unacceptably high. There have been numerous studies investigating injury characteristics and risk factors. Most studies use logistic regression models to analyze injury count data. However, logistic regression considers multiple injuries as an identical incident "having one or more injuries", therefore may undercount the total number of injuries and diminish the accuracy when examining injury risk factors. The present study is the first to identify the best count data model for injury data in the National Health Interview Survey (NHIS) and compare it to traditional logistic regression model.

4.1 Best Count Data Model for Injury Data in the National Health Interview Survey (NHIS)

Six count data models (Poisson, negative binomial (NB), hurdle Poisson (HP), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB)) were fitted to injury count data in the U.S. National Health Interview Survey. Hurdle negative binomial (HNB) model could not converge. This likely was attributable to a conflict between the model assumptions and the nature of our data. Although injury count data in NHIS are over-dispersed, non-zero count data may be under-dispersed. A further analysis on the distribution of non-zero count data showed that the mean of non-zero count (1.07) is approximately 10 times the variance (0.11), indicating under-dispersion. The zero-inflated negative binomial (ZINB) model turned out to be the best count data model. This finding is consistent with a recent study on modeling count data of Activities of Daily Living (ADL-s) with excess zeros [20]. However, Khan et al. [8] found that the HNB model offered the best fit to falls count data with excess zeros. In comparison, they analyzed falls count data from a prospective cohort study with 465 women aged 40-80 years old; we studied non-occupational injury count data from a national health survey with 248,850 participants aged 18-64 years old. Their data had 71% zero counts and 6% more than two counts; our data had 98.37% zero counts and 0.01% more than two counts. They examined the appropriateness of count models with only intercept; we compared count models with covariates. They also asserted that HNB might not offer the best fit for other injury count data and that the theoretical interpretations of models dealing with excess zeros could be improved.

The suitability of ZINB to our data may be attributable to the nature of NHIS survey design and the distribution of its injury count data. The NHIS questionnaire asked for the number of injuries that occurred during the last three months prior to the interview. This indicates that most people may not report any injuries (98.37% has no injuries), indicating excess zeros. Therefore, count data models specifically developed for handling excess zeros should offer a good fit to our data. Both zero-inflated and hurdle models can deal with excess zeros but have one important distinction in modeling zero counts. Zero-inflated models assume zero counts come from two groups: one group contains only zero counts (always-zero group), the other includes both zero and non-zero counts (not-always-zero group). In contrast, hurdle models assume that all zero counts come from one source, while all positive counts come from another source. In fact, zero-inflated models performed better than hurdle models in this study. This supports the theoretical interpretation that some people always score zero because they were in the low-risk group; others in the high- risk group might also score zero because they did not have any injuries during the three-month recall period (a relatively short

period for incurring injuries). In addition, ZINB model performed better than ZIP model due to previously mentioned over-dispersion of the NHIS data.

4.2 Comparison of ZINB Model and Logistic Regression Model for Analyzing Injury Data in NHIS

The ZINB model can predict the probability of a person being in the low-risk group and the expected number of injuries given he/she is in the high-risk group, and can examine injury risk factors in terms of injury frequency. Although the logistic regression model offered a good fit to our data, it examines injury risk factors in terms of the presence or absence of injuries rather than injury frequency. This distinction in examining injury risk factors between ZINB model and logistic regression model may lead to different conclusions. For example, a study of injury data in NHIS using logistic regression indicated that workers with disabilities were more likely to have non-occupational injuries than workers without disabilities [2]. However, our results showed that adults with disabilities and adults without disabilities have the same probability of being in the low-risk group; if in the high-risk group, adults with disabilities sustained approximately three times as many non-occupational injuries as their non-disabled counterparts.

4.3 Limitations of Injury Count Data in NHIS

Some limitations of our data should be noted. First, some independent variables are subject to moderate amounts of missing values. For example, 16.9 percent of the subjects have an unknown poverty status. This is perhaps because affluent families may be more reluctant to report their incomes. Although the missing data may result in bias in coefficient estimates, such bias is likely consistent over the years and therefore likely would not affect our conclusions. Another limitation of our data is that the NHIS only reports injuries that occurred during the last three months prior to the interview. Three-month period may be too short to detect any difference in the probability of being in the low-risk group, especially when comparing adults with and without disabilities.

5. CONCLUSION

The zero-inflated negative binomial (ZINB) model resulted in the best count data model for injury data in the National Health Interview Survey (NHIS) and demonstrated the potential to be the best model for injury count data with excess zeros. Given the infrequent occurrence of multiple medically consulted non-occupational injuries in NHIS, the logistic regression model is appropriate for statistical analysis. However, for more frequently-occurring injuries (e.g. sports injuries), logistic regression may undercount the total number of injuries and result in biased estimates since multiple injuries are collapsed into a single unit. The evaluation procedure and model selection criteria presented in this paper provide a useful approach to modeling injury count data with excess zeros.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Asada T, Kariya T, Kinoshita T, Asaka A, Morikawa S, Yoshioka M, et al. Predictors of fall-related injuries among community-dwelling elderly people with dementia. Age Ageing. 1996;25(1):22-8.

2. Price J, Shi J, Lu B, Smith GA, Stallones L, Wheeler KK, et al. Nonoccupational and occupational injuries to US workers with disabilities. Am J Public Health. 2012;102(9):38-46.
3. Smith AM, Stuart MJ, Wiese-Bjornstal DM, Gunnon C. Predictors of injury in ice hockey players. A multivariate, multidisciplinary approach. Am J Sports Med. 1997;25(4):500-7.
4. Zwerling C, Sprince NL, Wallace RB, Davis CS, Whitten PS, Heeringa SG. Risk factors for occupational injuries among older workers: An analysis of the health and retirement study. Am J Public Health. 1996;86(9):1306-9.
5. Liu W, Cela J. Count data models in SAS. In SAS Global Forum. 2008;317:1-12.
6. Karazsia BT, Van Dulmen MH. Regression models for count data: Illustrations using longitudinal predictors of childhood injury. J Pediatr Psychol. 2008;33(10):1076-1084.
7. Ullah S, Finch CF, Day L. Statistical modelling for falls count data. Accid Anal Prev. 2010;42(2):384-92.
8. Khan A, Ullah S, Nitz J. Statistical modelling of falls count data with excess zeros. Inj Prev. 2011;17(4):266-70.
9. Dunne RG, Asher KN, Rivara FP. Injuries in young people with developmental disabilities: comparative investigation from the 1988 National Health Interview Survey. Ment Retard. 1993;31(2):83-8.
10. Lombardi DA, Folkard S, Willetts JL, Smith GS. Daily sleep, weekly working hours, and risk of work-related injury: US National Health Interview Survey (2004-2008). Chronobiol Int. 2010;27(5):1013-1030.
11. Chen LH, Warner M, Fingerhut L, Makuc D. Injury episodes and circumstances: National Health Interview Survey, 1997-2007. Vital and health statistics. Series 10, Data from the National Health Survey. 2009;(241):1-55.
12. Landen DD, Hendricks SA. Estimates from the National Health Interview Survey on occupational injury among older workers in the United States. Scand J Work Environ Health. 1992;2:18-20.
13. Tiesman H, Zwerling C, Peek-Asa C, Sprince N, Cavanaugh JE. Non-fatal injuries among urban and rural residents: The National Health Interview Survey, 1997–2001. Inj Prev. 2007;13(2):115-119.
14. National Health Interview Survey (NHIS) Homepage. Accessed 29 December 2013. Available: http://www.cdc.gov/nchs/nhis.htm.
15. Cameron AC, Trivedi PK. Count data models for financial data. Statistical Methods in Finance. Handbook of Statistics. 1996;14:363-92.
16. Greene WH. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. NYU Working Paper; 1994. No.EC-94-10.
17. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989;57(2):307-33.
18. Hosmer DW, Lemeshow S. Applied logistic regression. Wiley.com; 2000.
19. Hu MC, Pavlicova M, Nunes EV. Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. Am J Drug Alcohol Abuse. 2011;37(5):367-75.
20. Zaninotto P, Falaschetti E. Comparison of methods for modelling a count outcome with excess zeros: application to Activities of Daily Living (ADL-s). J Epidemiol Community Health. 2011;65(3):205-10.

APPENDIX

1. COUNT DATA MODELS

Count data models can be used for injury frequency analysis because the number of injuries is a non-negative integer. This study compared six commonly used count data models: Poisson, negative binomial (NB), hurdle Poisson (HP), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) [1]. For the basic Poisson model, the probability of subject i having y_i injuries is

$$P(y_i) = \frac{\exp(-\mu_i)\cdot\mu_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, ..., \tag{1}$$

where μ_i is subject i's expected number of injuries. Poisson model specifies μ_i by using a log-linear function:

$$\mu_i = \exp(X_i'\beta), \tag{2}$$

Where $\mathbf{X}_i$ is a column vector of independent variables and $\boldsymbol{\beta}$ is a column vector of estimated coefficients [1,4].

However, the equi-dispersion (equal mean and variance) property of Poisson model restricts its applications in real-life data. Frequently, data are "over-dispersed" as the variance often exceeds the mean. This leads to underestimation of the standard errors of coefficient estimates and therefore incorrect inferences could be drawn. To account for over-dispersion, negative binomial (NB) model has been developed by introducing a new parameter λ_i:

$$\lambda_i = \exp(X_i'\beta + \varepsilon_i) = \mu_i \exp(\varepsilon_i), \tag{3}$$

where $\exp(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance $\boldsymbol{\alpha}$. The addition of this term allows the variance to exceed the mean as $V[y_i] = E[y_i] + \alpha \cdot E[y_i]^2$. The negative binomial probability mass function has the form:

$$P(y_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, \alpha > 0, \tag{4}$$

where $\boldsymbol{\Gamma}(\cdot)$ is a gamma function, $\mu_i = \exp(X_i'\beta)$ is subject i's expected number of injuries and α is the over-dispersion parameter. The negative binomial model reduces to the Poisson model as $\boldsymbol{\alpha}$ approaches 0. Therefore, when $\boldsymbol{\alpha}$ is significantly different from 0, negative binomial model should be applied; otherwise, Poisson model is preferred [1,4].

In addition to the over-dispersion, count data often have more observed zeros than expected from the Poisson model. This issue is known as "excess zeros". To account for excess zeros, hurdle models have been developed by assuming that zero counts and positive counts originate from two distinct data generating processes. For example, if the count of cups or glasses of alcohol consumed during the last year is the outcome, non-drinkers score zero

because they do not consume alcohol at all; while drinkers score positives as they should have consumed at least one glass of alcohol during the last year. Hurdle Poisson (HP) model can be derived by modeling positive counts with a zero truncated Poisson model. Its probability mass function has the form:

$$P(y_i) = \begin{cases} \theta_i & \text{if } y_i = 0 \\ \frac{(1-\theta_i)\cdot\exp(-\mu_i)\cdot\mu_i^{y_i}}{(1-\exp(-\mu_i))\cdot y_i!} & \text{if } y_i > 0 \end{cases}, \quad (5)$$

where θ_i is the probability of subject i having no injuries, $\mu_i = \exp(X_i'\beta)$ is subject i's expected number of injuries given subject i has at least one injury [1,2].

Hurdle Poisson (HP) model specifies the probability of subject i having no injuries (θ_i) by using a logit function:

$$\text{Log}\left(\frac{\theta_i}{1-\theta_i}\right) = Z_i'\gamma, \quad (6)$$

Where $\mathbf{Z}_i$ is a column vector of independent variables, γ is a column vector of estimated coefficients.

Similarly, hurdle negative binomial (HNB) model can be derived by modeling positive counts with a zero truncated negative binomial model. Its probability mass function has the form:

$$P(y_i) = \begin{cases} \theta_i & \text{if } y_i = 0 \\ \frac{(1-\theta_i)}{1-\left(\alpha^{-1}/(\alpha^{-1}+\mu_i)\right)^{\alpha^{-1}}} \cdot \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} & \text{if } y_i > 0 \end{cases}, \quad (7)$$

where $\mu_i = \exp\left(X_i'\beta\right)$ is subject i's expected number of injuries given subject i has at least one injury, α is the over-dispersion parameter and θ_i is the probability of subject i having no injuries; θ_i is specified by (6).

Zero-inflated models have also been developed to deal with excess zeros but its interpretation of zero counts is different from hurdle models. Instead of assuming all zero counts come from the same origin, zero-inflated models assume that zero counts are generated from two distinct groups: the always-zero group contains only zero counts and the not-always-zero group includes both zero and non-zero counts. For example, if the number of occupational injuries occurred during the last month is the outcome, some people can only score zero because they did not work in the last month. Other people who did work during the last month may also score zero because they were not injured at work. Therefore, some zeros come from people who can only score zeros (always-zero group); the other zeros are reported by people who sometimes score zeros (not-always-zero group).

Zero-inflated Poisson (ZIP) model can be derived by modeling positive counts in the not-always-zero group with a Poisson model. Its probability mass function has the form:

$$P(y_i) = \begin{cases} p_i + (1-p_i)\cdot \exp(-\mu_i) & \text{for } y_i = 0 \\ (1-p_i)\frac{\exp(-\mu_i)\cdot \mu_i^{y_i}}{y_i!} & \text{for } y_i > 0 \end{cases}, \quad (8)$$

where $\mu_i = \exp(X_i'\beta)$ is subject i's expected number of injuries given subject i is in the not-always-zero group and p_i is the probability of subject i from the always-zero group. Specification of p_i is done by a logit function

$$\text{Log}\left(\frac{p_i}{1-p_i}\right) = Z_i'\gamma, \quad (9)$$

Where $\mathbf{Z}_i$ is a column vector of independent variables, γ is a column vector of estimated coefficients [1,3].

Similarly, zero-inflated negative binomial (ZINB) model can be derived by modeling the not-always-zero group with a negative binomial model. Its probability mass function has the form:

$$P(y_i) = \begin{cases} p_i + (1-p_i)\cdot\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} & \text{for } y_i = 0 \\ (1-p_i)\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} & \text{for } y_i > 0 \end{cases}, \quad (10)$$

where $\mu_i = \exp(X_i'\beta)$ is subject i's expected number of injuries given subject i is in the not-always-zero group, α is the over-dispersion parameter and p_i is the probability of subject i from the always-zero group; p_i is specified by (9).

2. FORMULAS FOR CALCULATING ODDS RATIOS FROM THE ZINB MODEL AND THE LOGISTIC REGRESSION MODEL

Consider a 30 year old White married person who is college educated, not poor, and has medical insurance.

If he/she is disabled, probability of getting injured, YD > 0:

$P(Y^D > 0) = 1 - P(Y^D = 0) = (1 - P^D)\cdot(1 - g(\beta_0 + \beta_1))$, where $g(x)=\left(\frac{\alpha^{-1}}{\alpha^{-1}+e^x}\right)^{\alpha^{-1}}$.

Probability of not getting injured, Y^D =0:

$P(Y^D = 0) = P^D + (1 - P^D)\cdot g(\beta_0 + \beta_1)$

Odds of getting injured for a disabled person = $\frac{(1-P^D)\cdot(1-g(\beta_0+\beta_1))}{P^D+(1-P^D)\cdot g(\beta_0+\beta_1)}$,

where $\log\left(\frac{P^D}{1-P^D}\right) = Z_i'\gamma$.

If he/she is not disabled, probability of getting injured, YND>0:

$$P(Y^{ND} > 0) = 1 - P(Y^{ND} = 0) = (1 - P^{ND}) \cdot (1 - g(\beta_0))$$

Probability of not getting injured, YND=0:

$$P(Y^{ND} = 0) = P^{ND} + (1 - P^{ND}) \cdot g(\beta_0)$$

Odds of getting injured for a non-disabled person = $\frac{(1-P^{ND})\cdot(1-g(\beta_0))}{P^{ND}+(1-P^{ND})\cdot g(\beta_0)}$,

where $\log\left(\frac{P^{ND}}{1-P^{ND}}\right) = Z_i'\gamma$.

Based on the estimated coefficients of ZINB model (see Appendix Table 1),
$P^D = P^{ND} = 0.67764934$, $g(\beta_0 + \beta_1)$= 0.89209405, $g(\beta_0) = 0.96210191$
Odds ratio of getting injured between disabled people and non-disabled people = $\frac{\frac{(1-P^D)\cdot(1-g(\beta_0+\beta_1))}{P^D+(1-P^D)\cdot g(\beta_0+\beta_1)}}{\frac{(1-P^{ND})\cdot(1-g(\beta_0))}{P^{ND}+(1-P^{ND})\cdot g(\beta_0)}}$ =

$$\frac{\frac{1-g(\beta_0+\beta_1)}{P^D+(1-P^D)\cdot g(\beta_0+\beta_1)}}{\frac{1-g(\beta_0)}{P^{ND}+(1-P^{ND})\cdot g(\beta_0)}} = \frac{\frac{1-0.89209405}{0.67764934+(1-0.67764934)\cdot 0.89209405}}{\frac{1-0.96210191}{0.67764934+(1-0.67764934)\cdot 0.96210191}} = 2.91.$$

Based on the estimated coefficients of logistic regression model (see Appendix Table 2), odds ratio of getting injured between disabled people and non-disabled people = exp (1.10) = 3.

Appendix Table 1. Estimated coefficients for the ZINB model with a single covariate disability in the log linear part (P-value in parentheses)

Variable	ZINB model coefficient estimates			
	Logit part[1]		Log linear part[2]	
Constant	1.33*	(.00)	-3.23*	(.00)
Age group 2 indicator variable (1 if aged at 30-54, 0 otherwise)	0.46*	(.00)	--	--
Age group 3 indicator variable (1 if aged at 55-64, 0 otherwise)	0.41*	(.00)	--	--
Marital status 2 indicator variable (1 if single/never married, 0 otherwise)	-0.42*	(.00)	--	--
Marital status 3 indicator variable (1 if separated/divorced/widowed, 0 otherwise)	-0.75*	(.00)	--	--
Race 2 indicator variable (1 if non-Hispanic White, 0 otherwise)	-0.96*	(.00)	--	--
Race 3 indicator variable (1 if non-Hispanic Black, 0 otherwise)	-0.47*	(.00)	--	--
Race 4 indicator variable (1 if other races, 0 otherwise)	-0.71*	(.03)	--	--
College education indicator variable (1 if received college education, 0 otherwise)	-0.43*	(.00)	--	--
Poverty status 2 indicator variable (1 if near poor, 0 otherwise)	0.38*	(.00)	--	--
Poverty status 3 indicator variable (1 if not poor, 0 otherwise)	0.34*	(.00)	--	--
Poverty status 4 indicator variable (1 if unknown, 0 otherwise)	0.83*	(.00)	--	--
Disability status indicator variable (1 if disabled, 0 otherwise)	--	--	1.13*	(.00)
Medical insurance coverage status (1 if covered by medical insurance, 0 otherwise)	-0.46*	(.00)	0.30*	(.00)

**indicates significant estimated coefficients (P<.05). ZINB, zero-inflated negative binomial; NHIS, National Health Interview Survey*
--indicates not applicable.

1. $Log\left(\frac{p_i}{1-p_i}\right) = Z_i'\gamma.$

2. $\mu_i = exp(X_i'\beta).$

Appendix Table 2. Estimated coefficients for the logistic regression model with a single covariate disability (P-value in parentheses)

Variable	Logistic regression coefficient estimates	
Constant	-4.21*	(.00)
Disability status indicator variable (1 if disabled, 0 otherwise)	1.10*	(.00)

**indicates significant estimated coefficients (P<.05). NHIS, National Health Interview Survey.*

APPENDIX REFERENCES

1. Liu W, Cela J. Count data models in SAS. In SAS Global Forum. 2008;317:1-12.
2. Mullahy J. Specification and testing of some modified count data models. J Econometrics. 1986;33(3):341-65.
3. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1):1-14.
4. Cameron AC, Trivedi PK. Essentials of count data regression. A companion to theoretical econometrics. 2001;331-48. Blackwell Publishing Ltd.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=594&id=22&aid=5348