STATISTICAL MODELLING FOR FALLS COUNT DATA

Shahid Ullah[1], Caroline F Finch[1], Lesley Day[2]

[1]School of Human Movement and Sport Sciences, University of Ballarat, VIC 3353, Australia

[2]Accident Research Centre, Monash University, VIC 3800, Australia

Address for correspondence:

Dr Shahid Ullah

School of Human Movement and Sport Sciences

University of Ballarat

Mt Helen, Victoria, 3353, Australia

Ph: +61 3 5327 9061, Fax: +61 3 5327 9478

Email: s.ullah@ballarat.edu.au

ABSTRACT

Falls and their injury outcomes have count distributions that are highly skewed toward the right with clumping at zero, posing analytical challenges. Different modeling approaches have been used in the published literature to describe falls count distributions, often without consideration of the underlying statistical and modeling assumptions. This paper compares the use of modified Poisson and negative binomial (NB) models as alternatives to Poisson (P) regression, for the analysis of fall outcome counts. Four different count-based regression models (P, NB, zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB)) were each individually fitted to four separate fall count datasets from Australia, New Zealand and United States. The finite mixtures of P and NB regression models were also compared to the standard NB model. Both analytical (F, Vuong and bootstrap tests) and graphical approaches were used to select and compare models. Simulation studies assessed the size and power of each model fit. This study confirms that falls count distributions are over-dispersed, but not dispersed due to excess zero counts or heterogeneous population. Accordingly, the P model generally provided the poorest fit to all datasets. The fit improved significantly with NB and both zero-inflated models. The fit was also improved with the NB model, compared to finite mixtures of both P and NB regression models. Although there was little difference in fit between NB and ZINB models, in the interests of parsimony it is recommended that future studies involving modeling of falls count data routinely use the NB models in preference to the P or ZINB or finite mixture distribution. The fact that these conclusions apply across four separate datasets from four different samples of older people participating in studies of different methodology, adds strength to this general guiding principle.

Keywords: Fall count data, regression modeling, model fit, simulation study, negative binomial, zero-inflated models, finite mixture models

1. Introduction

Falls can have common and serious consequences for older people (Robertson et al., 2005). With an ageing population, the rise in the number of falls and the cost of their treatment is predicted to lead to a huge burden on the individual and the community (Moller, 2005). Falls epidemiology data describing the magnitude of, and trends in, the problem has largely been descriptive in nature (Boufous et al., 2006; Boufous et al., 2004). It is important that good statistical models are used to generate accurate and reliable information to guide policy decisions in relation to priority setting and intervention investments to tackle the fall injury problem. As with other areas of public health, there has been an increased interest in statistical modeling of injury count data, including falls outcomes, in recent years (Chin and Quddus, 2003; Lord et al., 2004; Lord et al., 2005; Robertson et al., 2005).

Datasets of the number of fall and fall-related injury have the form of discrete count data characterized by a large proportion of zero counts, with the remaining values being highly skewed toward the right. This is because fall incidents are relatively rare and most people will not sustain a serious injury if do they fall. Moreover, falls can also be recurrent events, in that over a period of time an individual may experience one or more falls (Williamson et al., 1996; Stalenhoef et al., 2002), and this recurrence aspect needs to be incorporated into appropriate statistical models of fall counts. In a very recent systematic review (Donaldson et al., 2009), fewer than one-third of the 83 reviewed papers used appropriate statistical methods to analyse falls as a recurrent event.

To further progress falls epidemiology, there is a need for a unified and justified approach to the use of appropriate statistical models for these data, taking into account the large proportion of zero counts and the possibility of recurrent falls. A number of published studies have incorrectly assumed a normal distribution when modelling falls count data and used Student's t test, linear regression, or analysis of variance, as has been highlighted elsewhere (Robertson et al., 2005). Other analysts have argued that falls count data does not meet the

usual normality assumption required of many standard statistical tests and have therefore relied on a transformation to induce normality (Slymen et al., 2006). This can be problematic in that transformations often do not yield normally distributed data and can make the interpretation of regression coefficients difficult because they are not estimated on the original scale (Byers et al., 2003).

An alternative, more common, approach has been to assume a Poisson (P) model which is better suited to fall count processes and has become quite widespread in public health to model the number of events or rates (Mwalili et al., 2008), especially when there are few incidents and hence, many observed zeros (Shankar et al., 1997). However, if the number of observed zeros far exceeds the expected number of zeros (equivalent to requiring that the mean is equal to the variance), then one of the key features of the P structure is violated. Often, falls count data exhibit more variability than the nominal variance under the P model, a condition called over-dispersion (in that the sample variance exceeds the mean). Such over-dispersion in count data can occur because of excess zeros, unexplained heterogeneity, or temporal dependency (Cameron and Trivedi, 1998). With regards to recurrent events, the P model assumes that such events occur independently of each other. This assumption is violated for fall outcomes, as a major risk factor for a subsequent fall is a previous fall (Donaldson et al., 2009; Hill et al., 1999).

The negative binomial (NB) model has a built-in dispersion parameter that can account for situations where the variance is greater than the mean (Chin and Quddus, 2003). A number of studies have therefore argued for the NB model as an alternative to the P model when count data are over-dispersed in relation to the mean (Bliss and Fisher, 2003; Byers et al., 2003; White and Bennetts, 1996). Such a modeling approach can also be appropriate when count data are recurrent (Glynn and Buring, 1996). The NB model explicitly accounts for the heterogeneity by modeling the Poisson mean as a Gamma random variable and introducing an extra dispersion parameter (Johnson et al., 2005; Lord, 2006).

Although P and NB models have been the most common choices to date, it is possible that they could still fail to fit a set of data with a lot of zeros because of zero-inflation, over-dispersion, or both (Deng and Paul, 2005). As an extension of standard P and NB models, zero-inflated count models have gained considerable recognition as an alternative means of handling count data with a preponderance of zeros (Lambert, 1992; Gupta et al., 1996; Li et al., 1999; Lord et al., 2004; Lord, 2006). For this type of count data, more zeros are observed than would be predicted by a normal P or NB process (Park and Lord, 2009; Lord et al., 2007; Warton, 2005). It is generally believed that data with excess zeros come from two sources or two distinct distributions, hence the apply-named dual state process. The underlying assumption of this two-state process gives a simple two-component mixture distribution with the first state having only zeros, while the other state leads to a standard P or NB count model. In general, the zeros from the first state are called structural zeros and those from the P or NB models are called sampling zeros or non-structural zeros.

In recent years, there has been considerable interest in regression models based on zero-inflated count models. Much of this interest stems from the seminal paper of Lambert (1992) though this type of model appears to have originated in the econometrics literature. Mullahy (1986) first formulated the zero-inflated Poisson (ZIP) regression model and such models have since been applied in many topic areas: the number of defects in a manufacturing process (Lambert, 1992); the abundance of rare species (Welsh et al. 1996); road accident frequencies (Shankar et al., 1997; Shankar et al., 2003; Qin et al., 2004; Kumara and Chin, 2003; Lee and Mannering, 2002); dental caries epidemiology (Bohning et al., 1999); pharmaceutical utilization and expenditure (Street et al., 1999); early growth and motor development (Cheung, 2002); and physical activity (Slymen et al., 2006).

In addition to zero-inflated models, there are many further extensions to the classical P and NB models, such as finite mixture models. These finite mixture models are particularly useful for heterogeneous populations that incorporate a combination of counts and continuous representation of population heterogeneity. For a mathematical derivation and

5

discussion of the application of finite mixture models, readers are referred to McLachlan and Peel (2000). Most recently, Park and Lord (2009) have proposed finite mixtures of P and NB models for analyzing motor vehicle crash data.

The modelling considerations raised above have significant implications for the description of falls data and published studies have used a variety of statistical approaches. To our knowledge, a full range of P and modified P (i.e. NB and zero-inflated) models have not been formally compared in terms of their applicability to falls data. Although Robertson et al. (2005) used the NB model in their consideration of statistical models for falls intervention trials, they compared it to two survival analysis models (the Andersen-Gill and marginal Cox regression) and not directly to other count distributions.

The aim of this paper is therefore compare the applicability of statistical count distributions to falls count data and to provide a clear rationale for future falls distribution-modeling approaches. In doing so, this study provides defensible guidance on how to appropriately model falls data in studies aiming to describe trends in injury numbers and rates. The paper has five objectives: to (1) overview the rationale for, and use of, P, NB, ZIP and zero-inflated negative binomial (ZINB) models; (2) to apply the four models to real-world falls count data and to compare how well the various models approximate this; (3) to formally compare the four models; (4) to report a statistical simulation experiment as a means of assessing the size and power of the model fit; and (5) to compare the NB model with finite mixtures of P or NB estimated using the same data.

2. Methods

A description of the data used in the example is first presented, so that the relevant features of the four regression models can be later described in the specific context of these data.

6

2.1 Description of falls data

Four sets of falls count data were obtained from published peer-reviewed studies and from personal communication with data owners. Two datasets were obtained from peer-reviewed randomized, controlled studies of falls in older people with published falls count data (Sydney, Australia: Voukelatos et al. (2007), Atlanta, United States: Wolf et al. (2003) respectively). A third dataset was obtained from a published prospective cohort study (Melbourne, Australia: Hill et al. (1999)). The fourth dataset from the Otago Exercise Programme in New Zealand was provided by the first author (Clare Robertson, personal communication) (Robertson et al., 2001). These four datasets were chosen to compare the modelling fits across a variety of data, generated from slightly different samples of older people from three different countries. We only used falls count data from the control groups of the randomized trials, to ensure consistency across all datasets.

Table 1 presents the four fall count distributions used in this study and shows the number of observed falls in the relevant study samples. In each study, over dispersion occurred with the variances exceeded the mean number of falls by a large amount. However, the improvement of the mean function could eliminate the over-dispersion (Miaou and Song 2005). Also the proportion of zeros does not appear to be over-represented given the sample mean of the data (Zorn 1998). It can be happen for falls count data where the number of falls may not exhibit excess zeros.

Table 1 also shows that the number of falls has a definite peak at no falls (i.e. zero counts) and decays rapidly after one or more falls. In particular, the structure of all distributions with a relatively high number of single falls and fewer very abundant falls is typical of all study datasets. Further examination of the data from the Sydney study, as an example, shows that it is over-dispersed, with mainly zero-counts (76% of the 337 study participants) and a sharp decay in the number of falls thereafter. The other datasets have a similar picture, but with a relatively lower frequency of no falls. Table 1 also shows the falls per 100 person years, which are estimated from the frequency of falls and follow up time. The Australian and New

Zealand falls rates are similar, ranging from 81 to 94 falls per 100 person years, compared to a higher rate among more frail older people in the Atlanta study.

Table 1: Distribution of falls count from four published studies

Recorded number of falls	Data source and study			
	Sydney Voukelatos et al. (2007)	Melbourne Hill et al. (1999)	New Zealand Robertson et al. (2001)	Atlanta Wolf et al. (2003)
0	256(76.0)*	49(51.0)*	64(53.3)*	56(39.7)*
1	54(16.0)	25(26.0)	33(27.5)	43(30.5)
2	14(4.2)	13(13.5)	16(13.3)	20(14.2)
3	10(3.0)	2(2.1)	3(2.5)	10(7.1)
4	1(0.3)	4(4.2)	1(0.8)	5(3.5)
5	2(0.6)	3(3.1)	0	5(3.5)
6	0	0	0	2(1.4)
7	0	0	0	0
8	0	0	1(0.8)	0
9	0	0	1(0.8)	0
10	0	0	1(0.8)	0
Total (number of people in each study)	337	96	120	141
Total number of falls	126	88	105	170
Total person years	155.5	96.0	111.8	130.2
Falls per person year	0.81	0.92	0.94	1.31
Mean	0.37	0.92	0.88	1.21
Variance	0.65	1.61	2.46	2.04
Sample characteristics	60+ years community dwelling relatively well	70+ years female community dwelling relatively well, no fall in previous 12 months	80+ years community dwelling relatively well	70+ years congregate living, transitionally frail

*** Data are the number (percent) of falls of this count in each individual dataset**

2.2 The Models

Poisson and negative binomial models

To set the scene for the subsequent development and comparison of models, it is necessary to introduce some notation here. Let $\mathbf{x} = \begin{bmatrix} x_1, \ldots\ldots\ldots, x_n \end{bmatrix}'$ denote the number of falls experienced by an individual within a given interval of time. The NB models the number of falls before a specified total number of falls is reached in a group of people. It can also be thought of as modelling the total number of people required before a specified number of falls is recorded or observed, thus motivating its name as the reversal of the binomial model. The NB model is commonly expressed in terms of its parameters μ and k such that

$$P(x_i = j) = \frac{\Gamma(j+k)}{\Gamma(j+1)\Gamma(k)} \left(\frac{\mu}{\mu+k} \right)^j \left(\frac{k}{\mu+k} \right)^k \quad \mu, k > 0, \ j = 0, \ 1, \ 2, \ldots\ldots$$

where μ is the average number of falls in a given time interval for each individual and k represents the degree of over-dispersion. The dispersion parameter represents the degree by which the variance of the number of falls differs from its mean. The variance of the NB model is $\sigma^2 = \mu(1 + \mu/k)$ and hence increasing values of k correspond to decreasing levels of dispersion. For $k \to \infty$, the model converges to the P model $P(x_i = j) = e^{-\mu}\mu^j / j!$ which has a null dispersion (the variance equals the mean). When both μ and k are positive (>0), the variance of the NB model is larger than the mean and the model can be applied to data with over dispersion.

Zero inflated models

The zero-inflated models are developed by adding a binary logit model as a mixing distribution in which two sets of data are combined in the proportions ϕ and $1 - \phi$, respectively. Let $\mathbf{y} = \begin{bmatrix} y_1, \ldots\ldots\ldots, y_n \end{bmatrix}'$ denote the number of falls, such that $y_i = 0$ with

probability ϕ and $y_i \sim P(x_i)$, i>0, with probability $1-\phi$. The mixed distribution of 'zero-falls' and 'one or more falls' becomes:

$$P(y_i = 0) = \phi + (1-\phi)P(x_i = 0)$$
$$P(y_i = j) = (1-\phi)P(x_i = j) \qquad j = 1, 2, ...$$

where $0 < \phi < 1$ is the probability of being an extra zero. For $\phi = 0$, these equations reduce to the ordinary P and NB models. If we consider the NB model, the expression of the ZINB model can be written as:

$$P(y_i = 0) = \phi + (1-\phi)\left(\frac{k}{\mu+k}\right)^k ; \quad \mu > 0; \ k > 0$$

$$P(y_i = j) = (1-\phi)\left(\frac{\Gamma(j+k)}{\Gamma(j+1)\Gamma(k)}\left(\frac{\mu}{\mu+k}\right)^j\left(\frac{k}{\mu+k}\right)^k\right); \quad \mu > 0; \ k > 0 \quad j = 1, 2, ...$$

The ZIP model is identical to ZINB but with a P probability.

$$P(y_i = 0) = \phi + (1-\phi)e^{-\mu}; \quad \mu > 0;$$

$$P(y_i = j) = (1-\phi)\frac{e^{-\mu}\mu^j}{j!}; \quad \mu > 0; \ j = 1, 2, ...$$

3. Model estimation framework

For each of the four model types, the maximum likelihood estimation (MLE) method was used to estimate μ, k and ϕ parameters and their corresponding standard errors and confidence limits for the falls count data, as relevant. The MLE was chosen, compared to other estimators, because it has properties of consistency, asymptotic normality and minimum variance for large samples. The MLE method was used to fit the falls data by applying a generalised linear model from underlying P, NB or zero-inflated families (Venables and Ripley, 2002). The models were fitted to the data using macros with the VGAM and PSCL packages written in R.

3.1 Model estimation results

Table 2 gives the values of the estimated parameters of the various modeled falls distributions. The smaller value of $\hat{k}$ for the Sydney study shows that the distribution is more dispersed in this study than in the others. A lower value of $\hat{\phi}$ indicates that the distributions are not zero-inflated when the ZINB model is fitted. Also none of the $\hat{\phi}$ parameters are significant for the ZINB, which indicates that the ZINB does not work. However, the data have excess zeroes in relation to fitting the ZIP model.

Table 2: Estimated parameters of the four fitted statistical models as fitted to falls count datasets. (95% confidence intervals are in parentheses.)

Data source and study	Model	$\hat{\mu}$	$\hat{k}$	$\hat{\phi}$
Sydney	P	0.37 (0.31-0.45)	-	-
Voukelatos et al.	NB	0.37 (0.30-0.47)	0.47 (0.22-0.72)	-
(2007)	ZIP	0.96 (0.73-1.26)	-	0.61 (0.51-0.70)
	ZINB	0.42 (0.06-3.18)	0.58 (0.00-4.08)	0.12 (0.02-0.95)
Melbourne	P	0.92 (0.74-1.13)	-	-
Hill et al. (1999)	NB	0.92 (0.70-1.21)	1.14 (0.25-2.02)	-
	ZIP	1.42 (1.07-1.88)	-	0.35 (0.22-0.52)
	ZINB	0.93 (0.39-2.22)	1.17 (0.00-3.41)	0.01 (0.00-1.00)
New Zealand	P	0.88 (0.72-1.06)	-	-
Robertson et al.	NB	0.88 (0.66-1.15)	0.80 (0.34-1.26)	-
(2001)	ZIP	1.42 (1.10-1.84)	-	0.39 (0.26-0.53)
	ZINB	0.88 (0.66-1.15)	0.80 (0.22-1.38)	0.01 (0.00-1.00)
Atlanta	P	1.21 (1.04-1.40)	-	-
Wolf et al. (2003)	NB	1.21 (0.99-1.47)	1.70 (0.60-2.80)	-
	ZIP	1.59 (1.31-1.94)	-	0.24 (0.15-0.38)
	ZINB	1.21 (0.99-1.47)	1.70 (1.05-2.35)	0.01 (0.00-1.00)

Note : P=Poisson model, NB=negative binomial model, ZIP=zero-inflated Poisson model,

ZINB= zero-inflated negative binomial model

The fitted models are also visualized in Figure 1. As expected, the zero counts are under-estimated and the single (one) fall count is over-estimated by fitting a P model. Overall, the NB and ZINB models provide the best fit for each dataset.

[Insert Figure 1 about here]

4. Model accuracy

The most common criterion for evaluating the performance of a statistical model is its accuracy in terms of fitting the data. Let f_i denote the observed frequency of ith fall and $\hat{f}_i$ denote the fitted frequency. The error is defined as $e_i = f_i - \hat{f}_i$ and the percentage error is $p_i = 100\, e_i / f_i$. Percentage errors have the advantage of being scale independent, so they are frequently used to compare model performance between different data series (Hyndman and Koehler, 2006). The most widely used measures of accuracy are:

1. The Mean Squared Error (MSE) $= mean\left(e_i^{2}\right)$, which is the average of all the squared errors ((Makridakis et al., 1998)

2. The Root Mean Squared Error (RMSE) $= \sqrt{MSE}$, which is often preferred to the MSE as it is on the same scale as the data

3. The Mean Absolute Error (MAE) $= = mean\left(|e_i|\right)$, which is less sensitive to extreme values than both the MSE and RMSE

4. The Mean Absolute Percentage Error (MAPE) $= mean\left(|p_i|\right)$, which is most commonly recommended because it has also many desirable properties including reliability, ease of interpretation, clarity of presentation, support of statistical evaluation, and the utilization of most of the information about the error (Hyndman and Koehler, 2006).

12

4.1 Model accuracy results

Table 3 shows the comparison of the fit of the four models to the published data according to these four accuracy measures. Smaller values of each of measures indicate more accurate, reliable and confident models. According to the MSE, RMSE and MAE results, the NB model provides the best fit for all four sets of falls data. However, there was little difference in fit between the ZINB and NB models. When considering the models with respect to MAPE, the NB and ZINB models both appeared to offer substantial improvements in fit over the standard P model, improvements not matched by the ZIP model.

Table 3: Mean square error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) of the models fitted to four fall distributions

Data source and study	Model	Model accuracy measures			
		MSE	RMSE	MAE	MAPE
Sydney	P	287.4	17.0	11.6	0.6
Voukelatos et al. (2007)	NB	4.7	2.2	1.7	0.4
	ZIP	21.1	4.6	3.3	0.4
	ZINB	5.1	2.3	1.8	0.4
Melbourne	P	41.8	6.5	5.4	0.7
Hill et al. (1999)	NB	3.2	1.8	1.3	0.5
	ZIP	8.7	2.9	2.5	0.7
	ZINB	3.2	1.8	1.3	0.5
New Zealand	P	41.4	6.4	4.2	0.6
Robertson et al. (2001)	NB	7.6	2.8	2.4	0.8
	ZIP	12.6	3.6	2.5	0.9
	ZINB	7.6	2.8	2.4	0.8
Atlanta	P	56.3	7.5	6.0	0.5
Wolf et al. (2003)	NB	3.4	1.8	1.6	0.2
	ZIP	23.2	4.8	3.7	0.4
	ZINB	3.4	1.9	1.6	0.2

Note : P=Poisson model, NB=negative binomial model, ZIP=zero-inflated Poisson model, ZINB= zero-inflated negative binomial model

5. Comparing models

Four criteria were used to compare and select among considered models: likelihood ratio, F test, Vuong statistic and bootstrap test. The likelihood ratio test is well understood and is not discussed further. The basic criterion of the F and bootstrap tests is to compare two models where one model should be nested with the other model (i.e. when one model is an

extension to the other). For example, the P model is nested within the NB model and there is therefore a need to test if there is over-dispersion due to heterogeneity. Another test is whether the dispersion parameter is necessary by comparing the P versus NB and ZIP versus ZINB models. In addition, as the P and NB models are nested within the ZIP and ZINB models, respectively, these models were tested for over dispersion due to excess zeros. Vuong's statistic has been used repeatedly as a goodness of fit measure to justify the selection of zero inflated models (Greene, 2000; Vuong, 1989).

5.1 The F-ratio test

Two nested models were considered. Using the notation of Shen and Faraway (2004), the large/full model is denoted by Ω, while the smaller/null model is denoted by ω and results from a linear restriction of the parameters of Ω. For example, the null model ω is

$$H_0 : P(x_i = j) = \frac{e^{-\mu} \mu^j}{j!}$$

and the full model Ω is

$$H_A : P(x_i = j) = \frac{\Gamma(j+k)}{\Gamma(j+1)\Gamma(k)} \left(\frac{\mu}{\mu+k} \right)^j \left(\frac{k}{\mu+k} \right)^k$$

The only difference between these two models is the non-zero dispersion parameter k.

The *F* statistic is used to test the hypothesis that the dispersion parameter k differs between two models.

$$F = \frac{(RSS_\omega - RSS_\Omega)/p-q}{RSS_\Omega/n-p} \sim F_{p-q,n-p}$$

where $RSS = \sum_{i=1}^{n} \left(f_i - \hat{f}_i \right)^2$ is the residual sum of squares; p = number of parameters in the full model; q = number of parameters in the null model and n = number of observations. In any particular case, it is necessary to determine which models represent the null and alternative hypotheses, to fit them and then compute the test statistic. However, The F test

cannot be used for testing the two models where the number of parameters is exactly the same.

5.2 The Vuong test

The Vuong test is based on the t statistic and has reasonable power in count data applications (Greene, 2000). The Vuong statistic (V-statistic) is computed as

$$V = \frac{\overline{m}\sqrt{N}}{S_m}$$

where $\overline{m}$ is the mean of $m = \ln\left[P_A(x_i = j)/P_0(x_i = j)\right]$ where $P_A(x_i = j)$ and $P_0(x_i = j)$ are the density functions of the full and null models and S_m and N are the standard deviation and sample size respectively (Vuong, 1989). For example, if to compare the ZIP and P models, P_A and P_0 are the density functions of the ZIP and P models respectively. The advantage of the Vuong test is that the entire distribution is used for comparison of the means, as opposed to just the excess zero mass (Shankar et al., 2003). A p-value less or equal to 0.05 for the Vuong statistic favors the full model, while a p-value greater than 0.05 indicates that the models are indistinguishable. The intuitive reasoning behind this test is that if the processes are not statistically different, then the mean ratio of their densities should equal one. The Vuong test can also be used to distinguish between the NB and ZIP, and NB and ZINB models as well (Shankar et al., 2003).

5.3 The bootstrap procedure

The idea behind bootstrapping a test statistic is to draw a large number (B) of random samples, which obey the null model (say P model) and, as far as possible, to re-assemble the real sample, and then compare the observed test statistic to the test statistics calculated from the bootstrap samples. The bootstrap p-values are estimated by the proportion of bootstrap statistics that are more extreme than the observed test statistic.

Although there are many ways to use the bootstrap for hypothesis testing, Walters and Campbell (2005) proposed a procedure for hypothesis testing to compute a bootstrap p value corresponding to the observed value of a test statistic T. In the falls count data, this study suggests a test procedure based on the following test statistic

$$T = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}_i^{\ w} - \hat{f}_i^{\ \Omega}\right]^2$$

where $\hat{f}_i^{\ w}$ and $\hat{f}_i^{\ \Omega}$ are the fitted frequencies of falls data under the null and full models. Two quantities are chosen to carry out the bootstrapping: a test statistic and a null model for the data under the null hypothesis. The *B* bootstrap values of the test statistic T_j^*, $j = 1,......,B$ are generated from the data under the null model. The achieved significance level, $p^*(T)$, is estimated by calculating the proportion of the bootstrap values which are $\geq T$, the value of the test statistic from the original falls data.

$$p^*(T) = P(T^* \geq T) = \frac{1}{B}\sum_{j=1}^{B}I\left(T_j^* \geq T\right)$$

The null model would be rejected if a small proportion of the T^* statistics are $\geq T$, yielding a small value of the $p^*(T)$.

5.4 Comparison of the fitted models

The p-values from the likelihood ratio, F, Vuong and bootstrap tests for comparing each pair of models are shown in Table 4. For example, comparing the NB and ZINB models for the Sydney data gives an F test with an estimated p value of 0.56, indicating that the two models are indistinguishable. In the interests of parsimony, the conclusion is that there is no need for the ZINB model with its extra parameter. Conversely, the P versus ZIP model comparisons all indicate that the four distributions have excess zeros. The reason for this finding is that the ZIP model can support a dual state process with excess zeros (Miaou, 1994). However, the ZIP versus ZINB and NB versus ZINB comparisons indicate that there is

no dispersion due to excess zeros. These excess zeros are more likely to arise when applying the NB model with small means than from the ZINB model (Warton, 2005). According to the Vuong test, the NB model was statistically valid ($p = 0.48$), compared to alternative forms including ZINB model. The higher p value for the NB and ZINB models in the other datasets shows that there is no difference between models of higher order. The F-ratio p values did not appear in Table 4 because of the same number of parameters for NB and ZIP models.

Table 4: Goodness of fit (likelihood ratio test (LRT), F-ratio (F), Vuong statistic (V) and bootstrap (BS)) p-values for comparing four fitted models to published falls count data

Models	Sydney Voukelatos et al. (2007)				Melbourne Hill et al. (1999)				New Zealand Robertson et al. (2001)				Atlanta Wolf et al. (2003)			
	LRT	F	V	BS	LRT	F	V	BS	LRT	F	V	BS	LRT	F	V	BS
P vs NB	<0.001	<0.001	-	<0.001	<0.001	<0.001	-	<0.001	<0.001	<0.001	-	<0.001	<0.001	<0.001	-	<0.001
P vs ZIP	<0.001	<0.001	0.01	<0.001	<0.001	<0.001	0.04	<0.001	<0.001	<0.001	0.01	<0.001	<0.001	<0.001	0.06	<0.001
P vs ZINB	<0.001	<0.001	-	<0.001	<0.001	<0.001	-	<0.001	<0.001	<0.001	-	<0.001	<0.001	<0.001	-	<0.001
NB vs ZIP*	0.88	-	0.15	0.44	0.76	-	0.12	0.45	0.52	-	0.10	0.62	0.35	-	0.33	0.47
NB vs ZINB	1.00	0.56	0.48	0.61	0.95	0.75	0.50	0.68	1.00	0.88	0.45	0.95	0.68	0.78	0.33	0.89
ZIP vs ZINB	0.04	<0.001	-	<0.001	0.01	<0.001	-	<0.001	<0.001	<0.001	-	<0.001	0.04	<0.001	-	<0.001

Note : P=Poisson model, NB=negative binomial model, ZIP=zero-inflated Poisson model, ZINB= zero-inflated negative binomial model

6. Simulation framework

Simulation studies are increasingly being used in the public health literature for a wide variety of situations (Vaeth and Skovlund, 2004). There are several advantages of simulations compared with collecting and/or analyzing real data (Burton et al., 2006; Demirtas, 2007). Firstly, a large number of samples of representative falls data can be created rather than being restricted to using only one (or just a few) dataset and this enables the distributions of statistical parameters to be estimated (the estimators). This is important because more reliable information would be obtained from a distribution of values than based on a single value. A second advantage of simulations enables the studying of phenomena that are too complex to describe and solve analytically.

6.1 Data generating process for size and power of the test

A bootstrap test was used to investigate the size and power within a simulation framework. The simulation approach provides estimates of statistical power and size that are derived from empirically derived distributions based on replications of simulated data generated from particular models of interest (Mooney, 1997; Wilkinson, 1999). Within the context of model selection, power and size estimates are based on the proportion of replications that indicate acceptable fit, with greater numbers of replications resulting in smaller confidence intervals (high power) around the estimates. With an existing asymptotic theory, the estimated power from the simulation was used to determine the performance of model selection criteria.

The idea behind the simulation approach for calculating size and power was derived from Davidson and MacKinnon (2006). Real data from the published falls distributions were fitted to the null and full models to obtain the estimated parameters. With application of the estimated parameters, simulated datasets produced from the null model can be used to calculate the size of the test and those corresponding to the full model used to calculate the power of the test. A bootstrapping test (as outlined in Section 5) was used to calculate the p-

values. The empirical power of a test was determined as the proportion of p-values from each simulated sample in which the null model is rejected at the actual significance level, when the null model is false. The size was calculated as the proportion of p-values from testing the null model on each simulated sample that are less than the actual significance level, when the null model is true. Simulations were run on a bi-processed Pentium 4 machine with a 3.20 GHz processor and 2.0 Gb RAM memory. The software used was R version 2.8.1 (R Development Core Team, 2004), operating on a Windows XP professional platform.

6.2 Simulation results

The initial models used to start the simulation process were taken to be the estimated models based on real falls data, to ensure realistic results. One hundred sets of falls data were simulated and the bootstrapping procedure was applied 100 times to each data set to obtain the significance levels of the test.

The rejection criteria of the bootstrap test for the four models at 1%, 5% and 10% significance levels are shown in Table 5. The bootstrap test performed satisfactorily for each model, although the actual size was slightly different from the simulated model size. For example, the actual size superseded the simulated size at 10% significance levels for all datasets, except for Atlanta, when the NB versus ZINB models were considered. In addition, there was a difference between actual and simulated sizes at the 5% significance level for the New Zealand data when the P and ZINB models were compared. Table 5 also shows the big gap between actual and estimated powers at 1%, 5% and 10% level of significance for all databases when the NB and ZINB models are considered.

21

Table 5: Simulated estimates (based on 100 simulation replications) of the size and power of the test to compare four models fitted to published falls count data

Data source and study	Comparison of models	Simulated model size $\Pr(P < \alpha) = \alpha$			Simulated model power $\Pr(P > \beta) = 1 - \beta$		
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\beta = 0.01$	$\beta = 0.05$	$\beta = 0.10$
Sydney	P vs NB	0.01	0.07	0.10	1.00	1.00	1.00
Voukelatos	P vs ZIP	0.01	0.04	0.09	1.00	1.00	1.00
et al. (2007)	P vs ZINB	0.00	0.02	0.05	1.00	1.00	1.00
	NB vs ZIP	0.00	0.00	0.00	1.00	0.98	0.94
	NB vs ZINB	0.01	0.06	0.12	1.00	0.34	0.39
	ZIP vs ZINB	0.01	0.02	0.04	0.48	0.66	0.74
Melbourne							
Hill et al.	P vs NB	0.01	0.04	0.08	0.88	0.98	0.98
(1999)	P vs ZIP	0.00	0.08	0.16	0.80	0.96	1.00
	P vs ZINB	0.02	0.04	0.12	0.76	0.90	0.94
	NB vs ZIP	0.00	0.03	0.06	0.08	0.28	0.41
	NB vs ZINB	0.02	0.10	0.14	0.00	0.10	0.14
New	ZIP vs ZINB	0.02	0.10	0.10	0.24	0.48	0.64
Zealand							
Robertson	P vs NB	0.04	0.06	0.08	1.00	1.00	1.00
et al. (2001)	P vs ZIP	0.00	0.02	0.02	0.90	1.00	1.00
	P vs ZINB	0.06	0.14	0.14	0.96	1.00	1.00
	NB vs ZIP	0.00	0.02	0.02	0.08	0.28	0.50
	NB vs ZINB	0.02	0.08	0.14	0.02	0.12	0.16
Atlanta	ZIP vs ZINB	0.00	0.02	0.06	0.38	0.54	0.70
Wolf et al.							
(2003)	P vs NB	0.02	0.02	0.06	0.96	0.96	1.00
	P vs ZIP	0.06	0.10	0.14	0.92	0.96	0.96
	P vs ZINB	0.00	0.02	0.05	1.00	1.00	1.00
	NB vs ZIP	0.00	0.03	0.07	0.84	0.56	0.42
	NB vs ZINB	0.01	0.03	0.12	0.00	0.06	0.14
	ZIP vs ZINB	0.08	0.08	0.08	0.92	0.92	1.00

Note : P=Poisson model, NB=negative binomial model, ZIP=zero-inflated Poisson model, ZINB= zero-inflated negative binomial model

7. Comparison of finite mixture models with standard and zero inflated P and NB models.

The Poisson and NB mixture models with a fixed number of components (K=2,3) were estimated with the expectation-maximization (EM) algorithm within a maximum likelihood framework and with Markov Chain Monte Carlo (MCMC) sampling within a Bayesian framework (Stasinopoulos and Rigby, 2007; Leisch, 2004). Models were compare using a penalised-likelihood approach for model selection: Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) (Park and Lord, 2009; Warton, 2005).

AIC and BIC values for comparing P and NB mixtures to the NB model are shown in Table 6. Based on these criteria, the NB model appears to be the better model than either the P and NB mixture models (K=2, K=3) for describing falls count data.

Table 6: Comparison of the negative binomial model against the Poisson and negative binomial mixture models: Log Likelihood (LL), Akaike's information criterion (AIC) and Bayesian information criterion (BIC) of the models selection to four fall distributions

Data source and study	Model	Model selection criteria		
		LL	AIC	BIC
Sydney	NB	- 267.9	539.9	547.5
Voukelatos et al. (2007)	FMP2	- 267.2	540.5	551.9
	FMP3	- 267.2	544.5	563.6
	FMNB2	- 267.9	545.8	564.9
	FMNB3	- 267.9	551.8	582.4
Melbourne	NB	- 127.3	258.6	263.7
Hill et al. (1999)	FMP2	- 126.8	259.7	267.3
	FMP3	- 126.8	263.6	276.5
	FMNB2	- 127.3	264.6	277.4
	FMNB3	- 127.1	270.2	290.7
New Zealand	NB	- 143.4	290.8	296.4
Robertson et al. (2001)	FMP2	- 143.1	292.3	300.6
	FMP3	- 142.5	295.0	308.9
	FMNB2	-143.5	297.1	311.0
	FMNB3	-143.5	303.1	325.3
Atlanta	NB	- 212.6	429.3	435.2
Wolf et al. (2003)	FMP2	- 211.7	429.5	438.3
	FMP3	- 211.7	433.5	448.2
	FMNB2	- 212.0	434.0	448.8
	FMNB3	- 212.6	441.3	464.9

Note: NB=negative binomial model, FMP2=Finite mixture Poisson (K=2), FMP3=Finite mixture Poisson (K=3), FMNB2=Finite mixture negative binomial (K=2), FMNB3=Finite mixture negative binomial (K=3)

8. Conclusions

There are several well-developed potential statistical models for analyzing falls count data but, to date, there has been little guidance on which is the most appropriate approach to use, and there are many published studies that have used incorrect statistical models for analyzing over-dispersion and recurrent fall events (Donaldson et al., 2009). Robertson et al. (2005) compared the NB model to two survival analysis models using two datasets, and concluded that the NB model was as appropriate as other statistical techniques (e.g. survival analysis methods), but was easier to use for analysis of falls trial data. Our study has further contributed to the field by directly comparing four different count-based regression models for analyzing falls data from four real-world studies conducted in Australia, New Zealand and the United States. In an attempt to identify the best models for analyzing falls count data, the four different counts were compared.

This study has shown that when applied to falls count data, use of the NB model gives better performance than the P, ZIP and ZINB models based on its superior statistical fit across all datasets. The NB model fit was also better than that achieved with finite mixture of P and NB regression models. Given the characteristics of the falls data, the study has also confirmed that falls count distributions are over-dispersed, but not dispersed due to excess zero counts or heterogeneous population. Although there was little difference in fit between NB and ZINB models, in the interests of parsimony it is recommended that future studies involving modeling of falls count data routinely use the NB models in preference to the others tested in this study. The fact that this conclusion holds true for four separate datasets from four different samples of older people participating in studies of different methodology, adds strength to this general guiding principle.

Although the NB model provided the best fit to the four sets of falls count data, this model may not necessarily be suitable for other count data. For example, the FMP and FMNB models might be useful where data are drawn from heterogeneous populations such as in crash data involving different vehicles at specified intersections (Park and Lord, 2009). In the

case of falls counts, the data are likely to be more homogenous because the falls occur in individual people and so finite mixture models are not so appropriate.

Acknowledgements

References

Bliss, C. I., Fisher, R. A., 2003. Fitting the negative binomial distribution to biological data. Biometrics 9, 176-200.

Bohning, D., Dietz, E., Schlattmann, P., 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society A 162, 195-209.

Boufous, S., Finch, C.F., Lord, S.R., 2004. Incidence of hip fracture in New South Wales: are our efforts having an effect? The Medical journal of Australia, 180, 623-626.

Boufous, S., Finch, C., Lord, S., Close, J., Gothelf, T., Walsh, W., 2006. The epidemiology of hospitalised wrist fractures in older people, New South Wales, Australia. Bone, 39, 1144-1148.

Burton, A., Altman, D. G., Royston, P., Holder, R. L., 2006. The design of simulation studies in medical statistics. Statistics in Medicine 25, 4279-4292.

Byers, A. L., Allore, H., Gill, T. M., Peduzzi, P. N., 2003. Application of negative binomial modelling for discrete outcomes: A case study in ageing research. Journal of Clinical Epidemiology 56, 559-564.

Cameron, A. C., Trivedi, P. K., 1998. Regression Analysis of Count Data. pp. 70-77. Cambridge University Press, UK.

Cheung, Y. B., 2002. Zero-inflated models for regression analysis of count data: A study of growth and development. Statistics in Medicine 21, 1461-1469.

Chin, H. C., Quddus, M. A., 2003. Modeling count data with excess zeroes: An empirical application to traffic accidents. Sociological Methods and Research 32, 90-116.

Davidson, R., MacKinnon, J. G., 2006. The power of bootstrap and asymptotic tests. Journal of Econometrics 133, 421-441.

Demirtas, H., 2007. Letter to the Editor re: The design of simulation studies in medical statistics. Statistics in Medicine 26, 3818-3821.

Deng, D., Paul, S. R., 2005. Score tests for zero-inflation and over-dispersion in generalized linear models. Statistica Sinica 15, 257-276.

Donaldson, M. G., Sobolev, B., Cook, W. L., Janssen, P. A., Khan, K. M., 2009. Analysis of recurrent events: A systematic review of randomised controlled trials of interventions to prevent falls. Age and Ageing 38, 151-155.

Glynn, R. J., Buring, J. E., 1996. Ways of measuring rates of recurrent events. British Medical Journal 312, 364-367.

Greene, W.H., 2000. Econometric Analysis, 4th ed. edn, Prentice Hall, New Jersey.

Gupta, P. L., Gupta, R. C., Tripathi, R. C., 1996. Analysis of zero-adjusted count data. Computational Statistics and Data Analysis 23, 207-218.

Hill, K., Schwarz, J., Flicker, L., Carroll, S., 1999. Falls among healthy, community-dwelling, older women: A prospective study of frequency, circumstances, consequences and prediction accuracy. Australian and New Zealand Journal of Public Health 23, 41-48.

Hyndman, R. J., Koehler, A. B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679-688.

Johnson, N. L., Kemp, A. W., Kotz, S., 2005. Univariate Discrete Distributions. pp 208-250 John Wiley, New York.

Kumara, S. S., Chin, H. C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 4, 53-57.

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

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis. Accident Analysis and Prevention 34, 149-161.

Leisch, F., 2004. FlexMix: A general framework for finite mixture models and latent class regression in R", Journal of Statistical Software 11, 1-18.

Li, C., Lu, J., Park, J., Kim, K., Brinkley, P. A., Peterson, J. P., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics 41, 29-38.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38, 751-766.

Lord, D., Washington, S. P., Ivan, J. N., 2004. Statistical challenges with modeling motor vehicle crashes: Understanding the implications of alternative approaches. Research report, Center for Transportation Safety, Texas Transportation Institute, Texas A&M University System, http://tti.tamu.edu/documents/CTS-04_01.pdf.

Lord, D., Washington, S. P., Ivan, J. N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accident Analysis and Prevention 37, 35-46.

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39, 53-57.

Makridakis, S., Wheelwright, S. C., Hyndman, R. J., 1998. Forecasting: Methods and Applications. pp. 42-43. John Wiley, New York.

McLachlan, G., Peel, D, 2000. Finite Mixture Models, John Wiley, New York.

Miaou, S. P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention *26*, 471-482.

Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis and Prevention 37, 699-720.

Moller, J., 2005. Current costing models: Are they suitable for allocating health resources? the example of fall injury prevention in Australia. Accident Analysis and Prevention 37, 25-33.

Mooney, C.Z., 1997. Monte Carlo Simulation. Sage University Press, Iowa City, IA.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics. 33, 341-365.

Mwalili, S. M., Lesaffre, E., Declerck, D., 2008. The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. Statistical Methods in Medical Research 17, 123-139.

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis", Accident Analysis and Prevention*, DOI:10.1016/j.aap.2009.03.007*

Qin, X., Ivan, J. N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36, 183-191.

R Development Core Team, 2004. A language and environment for statistical computing. R foundation for statistical computing: Vienna, Austria*,* http://www.R-project.org

Robertson, M. C., Campbell, A. J., Herbison, P., 2005. Statistical analysis of efficacy in falls prevention trials. The Journals of Gerontology 60, 530-534.

Robertson, M. C., Gardner, M. M., Devlin, N., McGee, R., Campbell, A. J., 2001. Effectiveness and economic evaluation of a nurse delivered home exercise programme to prevent falls. 2: Controlled trial in multiple centres. British Medical Journal 322, 701-704.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis and Prevention 29, 829-837.

Shankar, V. N., Ulfarsson, G. F., Pendyala, R. M., Nebergall, M. B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41, 627-640.

Shen, Q., Faraway, J., 2004. An F test for linear models with functional response. Statistica Sinica 14, 1239-1257.

Slymen, D., Ayala, G., Arredondo, E., Elder, J., 2006. A demonstration of modeling count data with an application to physical activity. Epidemiologic Perspectives and Innovations 3, 1-9.

Stalenhoef, P. A., Diederiks, J. P., Knottnerus, J. A., Kester, A.D., Crebolder, H.F., 2002. A risk model for the prediction of recurrent falls in community-dwelling elderly: a prospective cohort study. Journal of clinical epidemiology 55, 1088-094.

Stasinopoulos, D.M., Rigby, R.A. 2007. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23, 1-46.

Street, A., Jones, A., Furuta, A., 1999. Cost-sharing and pharmaceutical utilisation and expenditure in russia. Journal of Health Economics 18, 459-472.

Vaeth, M., Skovlund, E., 2004. A simple approach to power and sample size calculations in logistic regression and Cox regression models. Statistics in Medicine 23, 1781-1792.

Venables, W. N., Ripley, B. D., 2002. Modern applied statistics with S. pp. 23-25. Springer, Germany.

Voukelatos, M., Cumming, R. G., Lord, S. R., Rissel, C., 2007. A randomised, controlled trial of tai chi for the prevention of falls: The central Sydney tai chi trial. Journal of the American Geriatrics Society 55, 1185-1191.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* 57, 307-333.

Walters, S. J., Campbell, M. J., 2005. The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes. Statistics in Medicine 24, 1075-1102.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16, 275-289.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., Lindenmayer, D. B., 1996. Modelling the abundance of rare species: Statistical models for counts with extra zeros. Ecological Modelling 88, 297-308.

White, G. C., Bennetts, R. E., 1996. Analysis of frequency count data using the negative binomial distribution. Ecology 77, 2549-2557.

Wilkinson, L., Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 54, 594-604.

Williamson, D. S., Bangdiwala, S. I., Marshall, S. W., Waller, A. E., 1996. Repeated measures analysis of binary outcomes: Applications to injury research. Accident Analysis and Prevention 28, 571-579.

Wolf, S. L., Sattin, R. W., Kutner, M., O'Grady, M., Greenspan, A. I., Gregor, R. J., 2003. Intense tai chi exercise training and fall occurrences in older, transitionally frail adults: A randomized, controlled trial. Journal of the American Geriatrics Society 51, 1693-1701.

Zorn, C., 1998. Evaluating zero-inflated and hurdle Poisson specifications. Sociological

Methods and Research 26, 368-400.

FIGURE LEGENDS

Figure 1. Actual and fitted falls count data from published studies: (a) Sydney, Australia (Voukelatos et al., 2007); (b) Melbourne, Australia (Hill et al., 1999). (c) Otago, New Zealand (Robertson et al., 2001) and (d) Atlanta, United States (Wolf et al., 2003)