STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2907–2920 (DOI: 10.1002/sim.860)

Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme

Kelvin K. W. Yau[1] and Andy H. Lee[2,*,†]

[1] *Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong*
[2] *Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, GPO Box U 1987, Perth, WA 6845, Australia*

SUMMARY

This study presents a zero-inflated Poisson regression model with random effects to evaluate a manual handling injury prevention strategy trialled within the cleaning services department of a 600 bed public hospital between 1992 and 1995. The hospital had been experiencing high annual rates of compensable injuries of which over 60 per cent were attributed to manual handling. The strategy employed Workplace Risk Assessment Teams (WRATS) that utilized a workplace risk identification, assessment and control approach to manual handling injury hazard reduction. The WRATS programme was an intervention trial, covering the 1988–1995 financial years. In the course of compiling injury counts, it was found that the data exhibited an excess of zeros, in the context that the majority of cleaners did not suffer any injuries. This phenomenon is typical of data encountered in the occupational health discipline. We propose a zero-inflated random effects Poisson regression model to analyse such longitudinal count data with extra zeros. The WRATS intervention and other concomitant information on individual cleaners are considered as fixed effects in the model. The results provide statistical evidence showing the value of the WRATS programme. In addition, the methods can be applied to assess the effectiveness of intervention trials on populations at high risk of manual handling injury or indeed of injury from other hazards. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

The overall annual economic burden of serious occupational injuries in Australia has been estimated at $20 billion. However, the burden and type of injury is uneven across or within industries. For example, in 1992–1993 it was estimated that the 400 000 staff (8 per cent of the national work force) working in the health and community services sector sustained an injury rate 30 per cent higher than the all-industries rate [1]. Within hospitals, registered

*Correspondence to: Andy H. Lee, Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, GPO Box U 1987, Perth, WA 6845, Australia
†E-mail: Leea@health.curtin.edu.au

Contract/grant sponsor: Research Grant Council of Hong Kong

Received January 2000
Accepted September 2000

nurses accounted for 26 per cent of all injuries and cleaners 9.2 per cent, but when examined in terms of incidence rates, the rates for registered nurses were lower than the average for all occupations whereas the rates for cleaners were twice the average.

A 600 bed adult teaching hospital in Western Australia (the study hospital) was concerned because of its high injury rates and an associated decade of annually increasing workers' compensation premiums. For the 1990–1991 financial year, the hospital's rate of injury was approximately 25 per cent higher than the average for the State. Manual handling was the mechanism in 60 per cent of cases, compared with approximately 30 per cent for all industries in the state.

While it was clear that manual handling injury risk reduction was the corporate safety priority, evidence-based risk reduction protocols were not readily available. Many hospital employees still perceived the nature of their work and inherent risks as immutable. Previous research in ergonomics [2] indicated that manual handling training alone was ineffective in reducing risk. Within Western Australia, manual handling was the single most common mechanism of work related injury, and had showed no reduction for years.

Late in 1992 a pilot programme of manual handling risk assessment teams was introduced to the Cleaning Services of the study hospital by its Department of Occupational Health. The teams were established in response to a consultative style manual handling State regulations. The aim was to identify and assess manual handling risk and recommend controls to executive management. These 'Workplace Risk Assessment Teams', or WRATS (the acronym designed to give the impression of the teams scurrying around), had a membership that comprised safety and health representatives from within the organization, as well as ordinary cleaning staff. The cleaners were selected because they contributed the highest injury rates among all occupational groups in the hospital. In addition, their duties were relatively autonomous for implementation of WRATS.

Manual handling risk assessment, in consultation with employees, continues to be promoted in many jurisdictions worldwide, yet proper evaluation of its effectiveness is scarce. A study [3] implied employees may not be competent to prioritize manual handling injury risk, whilst another study [4] suggested that participatory ergonomics teams can be successful in identifying problems and implementing solutions. There is emerging evidence [5] that employee participation can reduce overall injury risk. For the study hospital, it was found that the teams were 'well received' and a number of workplace changes made, but no formal injury outcome evaluation was ever undertaken.

This paper sets out to evaluate the effectiveness of the WRATS intervention in reducing the number of manual handling injuries. Traditionally, the Poisson distribution is adopted to model injury counts in occupational health. An inspection of the record, however, reveals that zero constituted in excess of 65 per cent of the observations. Such extra zeros violate the implicit mean-variance relationship of the Poisson error structure. This phenomenon of *overdispersion* has generally been considered in the literature [6]. Specific models [7] for overdispersion (some incorporating random effects) arise from alternative possible mechanisms for the underlying process. In our situation, the extra-Poisson variation induced by the excess zeros can be accommodated through a compound probability model for the response, namely the zero-inflated Poisson distribution. Consequently, it is logical to propose a zero-inflated random effects Poisson regression model to analyse such longitudinal count data with extra zeros. Analysis was delayed for 5 years to allow for maturation of workers' compensation claims incurred during the study period.

2. DATA FROM AN OCCUPATIONAL INJURY PREVENTION PROGRAMME

The earliest date that pre-intervention data could be accurately collected was 1 July 1988 (commencement of the 1988/1989 financial year). The WRATS programme was officially implemented on 1 November 1992. The study period concluded on 31 October 1995 due to a number of industrial issues as the cleaning operation was under threat of being 'contracted out'.

2.1. Study group

In the study hospital, 507 people were ever employed in cleaning services. Staffing number remained relatively constant over the seven-year period. Accounting for employee turnover, we shall focus on the cohort of 137 cleaners who were present in both pre- and post-WRATS intervention. Of those employed, females comprised 77 per cent of the study group.

2.2. Data sources

Data (including hours worked) were obtained from fortnightly financial records, personnel records, incident data sheets, hospital workers' compensation files and workers' compensation records held by the insurer.

2.3. Variables

The outcome variable of interest is *injury count Y*. Measurement error is not a problem as only compensable injuries are used. These claims have been independently approved by the insurer. The extent of exposure is given by *hours worked*, defined to be the number of hours actually worked but does not include leave or overtime payment. At this stage severity status cannot be determined despite the fact that a serious injury may affect subsequent injury rate. To control for confounding factors influencing injury risk, covariates *age* and *gender* are included. Other potential confounders such as smoking, psychosocial factors, height and weight are not available. Educational background could also be a confounder, however, it is generally accepted that occupation itself is a good surrogate for educational achievement and this group of cleaners is categorized as 'labourers' under the Australian Bureau of Statistics classification.

Prima facie evidence based on the empirical distribution of injury counts (Table I) suggests the introduction of WRATS within Cleaning Services has led to a sustained reduction in manual handling injuries, yet the extent of individual exposure (hours worked by staff) and other concomitant information have not been taken into consideration. It is imperative to confirm the apparent injury reduction by an appropriate statistical model. From Table I, it

Table I. Injury frequency distribution pre- and post-WRATS.

Y	0	1	2	3	4	8
Pre	72	38	17	6	3	1
Post	108	24	5	0	0	0

can be seen that there is a high frequency at zero representing the injury-free cleaners. This pattern is typical of data in occupational health. If the Poisson assumption holds, the mean and variance should coincide. Böhning *et al.* [8] proposed an overdispersion test which is simple to apply. Here, the overdisperison test delivers the z-values 5.625 (p-value ≈ 0) and 0.442 (p-value $= 0.329$) for the pre-WRATS and post-WRATS counts, respectively, indicating a strong overdispersion for the pre-WRATS period. The homogeneous Poisson distribution does not provide an adequate fit to the pre-WRATS data. Indeed, the Pearson chi-square statistic has the value 6.63 based on two degrees of freedom.

3. ZERO-INFLATED POISSON MODELS

Zero-inflated Poisson (ZIP) regression is a model for count data with excess zeros. Consider a discrete random variable Y with ZIP distribution [9]

$$\Pr(Y=0) = \phi + (1-\phi)\,\mathrm{e}^{-\theta}$$

$$\Pr(Y=y) = (1-\phi)\frac{\mathrm{e}^{-\theta}(\theta)^y}{y!} \quad y = 1,2,\ldots$$

where $0<\phi<1$ so that it incorporates more zeros than those allowed by the Poisson. A graphical representation of this distribution is given by Böhning *et al.* [10]. The ZIP distribution may also be regarded as a mixture of a Poisson(θ) and a degenerate component putting all its mass at zero. A plausible interpretation in the context of the cleaners population is thus in terms of its (unobserved) two-point heterogeneity: a subpopulation of cleaners who are not at risk of manual handling injury, and another subpopulation whose members are susceptible and may incur injury several times. It can be shown that

$$E(Y) = (1-\phi)\theta$$

$$\mathrm{var}(Y) = E(Y) + E(Y)\{\theta - E(Y)\}$$

so that the ZIP model incorporates the extra variation unaccounted for by the Poisson. Further properties, including maximum likelihood inference, can be found in Gupta *et al.* [11].

For independent counts Y_i, $i=1,\ldots,n$, a ZIP regression model [12] was proposed to study the effects of risk factors or confounders by assuming both $\log(\theta_i)$ and $\mathrm{logit}(\phi_i) = \log(\phi_i/(1-\phi_i))$ to be linear functions of some covariates. Maximum likelihood estimation of the regression coefficients is performed via the EM algorithm. A score test for zero inflation is proposed by van den Broek [13].

Let $p = (1-\phi)(1-\mathrm{e}^{-\theta})$. The ZIP distribution can be equivalently expressed as

$$\Pr(Y=0) = 1-p$$

$$\Pr(Y=y) = p\,\frac{\mathrm{e}^{-\theta}\theta^y}{y!}\Big/(1-\mathrm{e}^{-\theta}) \quad y = 1,2,\ldots$$

Here p is the probability of observing at least one injury and, conditional on at least one injury, the injury count is described by the truncated Poisson distribution [14]. With this parameterization, a conditional ZIP regression was advocated [15], which is simpler to fit than Lambert's formulation [12]. Following such a conditional approach, we derive a random effects ZIP regression model to analyse longitudinal count data with extra zeros, taking into account the extent or time of individual exposures.

4. ZIP REGRESSION MODEL WITH RANDOM EFFECTS

4.1. Model and log-likelihood

The response variable is the injury count of individuals in pre-WRATS intervention and post-WRATS intervention. Let Y_{ij}, $i=1,2,\ldots,n$; $j=1,2$ be the injury count random variable of the ith individual in the jth period. It is reasonable to assume that the injury counts are independent between individuals. However, it is anticipated that, for the ith individual, observations Y_{i1} and Y_{i2} should exhibit certain within-person correlation. Such a correlation can be modelled explicitly by random effects attached to each individual using the method of generalized linear mixed models (GLMM) [16–19]. Following the conditional approach [15], a logistic component is applied to model the probability of injury incidence p_{ij}. Given that injury has ever occurred in the observed period, the mean number of injuries sustained may be modelled by a truncated Poisson (θ_{ij}) distribution.

In the regression setting, both $\text{logit}(p_{ij})$ and $\log(\theta_{ij})$ are assumed to depend on a linear function of covariates. The covariates appearing in these two parts are not necessarily the same. In addition, the time of exposure t_{ij} is also available for each Y_{ij}. In the context of generalized linear models, it is desirable to incorporate an offset term $\log t_{ij}$ to reduce error variation, giving

$$\text{logit}(p_{ij}) = \log t_{ij} + \xi_{ij}$$

$$\log(\theta_{ij}) = \log t_{ij} + \eta_{ij} \quad \text{for those observations with } y_{ij} > 0$$

With random effects, the linear predictors ξ_{ij} and η_{ij} are defined as follows:

$$\xi_{ij} = w_{ij}'\alpha + u_i$$

$$\eta_{ij} = x_{ij}'\beta + v_i$$

where w_{ij} and x_{ij} are, respectively, vectors of covariates for the logistic and the truncated Poisson components and α and β are vectors of regression coefficients. Letting $u=(u_1,\ldots,u_n)'$ and $v=(v_1,\ldots,v_m)'$, where m is the number of individuals with at least one non-zero injury count in the observed periods, we assume that u, v are independent and are distributed as $\text{N}(0,\sigma_u^2 I_n)$ and $\text{N}(0,\sigma_v^2 I_m)$, respectively, where I_p denotes a $p\times p$ identity matrix.

Following the GLMM method [18], the best linear unbiased prediction (BLUP) type log-likelihood is given by $l = l_1 + l_2$, where

$$l_1 = \sum_{y_{ij}=0} \log\left(\frac{1}{1+t_{ij}\exp\xi_{ij}}\right) + \sum_{y_{ij}>0} \log\left(\frac{t_{ij}\exp\xi_{ij}}{1+t_{ij}\exp\xi_{ij}}\right)$$
$$+ \sum_{y_{ij}>0} (y_{ij}(\log t_{ij} + \eta_{ij}) - t_{ij}\exp\eta_{ij} - \log(1 - e^{t_{ij}(-\exp\eta_{ij})}) - \log(y_{ij}!))$$
$$l_2 = -\tfrac{1}{2}[n\log(2\pi\sigma_u^2) + \sigma_u^{-2}u'u + m\log(2\pi\sigma_v^2) + \sigma_v^{-2}v'v]$$

This conditional model has the advantage of a separate parameterization, which is simple to interpret and readily fitted by a Newton–Raphson or quasi-Newton algorithm. With such a separate parameterization, it is then fully efficient to fit the components separately. Here the regression coefficients α are interpreted directly in terms of the probability of injury and the β coefficients relate separately to the mean count once injured.

4.2. *Estimation via Newton–Raphson algorithm*

After the incorporation of offsets and random effects, it is obvious that the log-likelihood of our model can also be separated into two components: one contains (α, u) and the other contains (β, v). In detail, the log-likelihood l may be rewritten as the sum of two separate components, l_ξ and l_η, where

$$l_\xi = \sum_{y_{ij}=0} \log\left(\frac{1}{1+t_{ij}\exp\xi_{ij}}\right) + \sum_{y_{ij}>0} \log\left(\frac{t_{ij}\exp\xi_{ij}}{1+t_{ij}\exp\xi_{ij}}\right)$$
$$-\tfrac{1}{2}[n\log(2\pi\sigma_u^2) + \sigma_u^{-2}u'u]$$
$$l_\eta = \sum_{y_{ij}>0} (y_{ij}(\log t_{ij} + \eta_{ij}) - t_{ij}\exp\eta_{ij} - \log(1 - e^{t_{ij}(-\exp\eta_{ij})}) - \log(y_{ij}!))$$
$$-\tfrac{1}{2}[m\log(2\pi\sigma_v^2) + \sigma_v^{-2}v'v]$$

Therefore, estimation can be performed by using the following two sets of Newton–Raphson algorithms according to the derivatives of l_ξ and l_η:

$$\begin{bmatrix}\tilde{\alpha}\\ \tilde{u}\end{bmatrix} = \begin{bmatrix}\alpha_0\\ u_0\end{bmatrix} + \Im_{\alpha,u}^{-1}\begin{bmatrix}\dfrac{\partial l_\xi}{\partial\alpha}\\ \dfrac{\partial l_\xi}{\partial u}\end{bmatrix}$$

$$\begin{bmatrix}\tilde{\beta}\\ \tilde{v}\end{bmatrix} = \begin{bmatrix}\beta_0\\ v_0\end{bmatrix} + \Im_{\beta,v}^{-1}\begin{bmatrix}\dfrac{\partial l_\eta}{\partial\beta}\\ \dfrac{\partial l_\eta}{\partial v}\end{bmatrix}$$

where $\alpha_0, u_0, \beta_0, v_0$ are initial values of α, u, β, v, respectively, and are replaced by the updated estimates in each iteration for given values of σ_u^2 and σ_v^2. $\Im_{\alpha,u}$ is the negative second derivative of l_ξ with respect to α and u. $\Im_{\beta,v}$ is the negative second derivative of l_n with respect to β and v. Expressions for the first and second derivatives of l_1 with respect to ξ and η are given in Appendix A.

4.3. Variance component estimation

In the Newton–Raphson algorithm, it is assumed that the variance components σ_u^2 and σ_v^2 are given. In practice, these are unknown and required to be estimated. The estimation procedure is essentially iterative. For given initial values of σ_u^2 and σ_v^2, the Newton–Raphson recursive equations are used to estimate the elements in the linear predictor. When convergence is attained, σ_u^2 and σ_v^2 are estimated by the most updated values of $\tilde{u}, \tilde{v}$ and the corresponding elements of the information matrices. The process continues until all parameter estimates converge. For the estimation of variance components, the residual maximum likelihood (REML) approach is usually adopted to correct the bias due to maximum likelihood estimation. Estimators of variance components, their asymptotic variances, and variances of the fixed effect parameters are given in Appendix B.

4.4. Assessing high risk individuals

In our model setting, $\tilde{u}$ and $\tilde{v}$ are estimated as predictions of the random effects in the respective logistic and truncated Poisson components. For diagnostic purpose, it may be useful to standardize the random effect predictions. In particular, the standardized random effect predictions are calculated as $\frac{\tilde{u}_k}{\sqrt{(\operatorname{var}\tilde{u}_k)}}$ in the logistic component and as $\frac{\tilde{v}_l}{\sqrt{(\operatorname{var}\tilde{v}_l)}}$ in the truncated Poisson component, where $\operatorname{var}\tilde{u}_k$ is the kth diagonal entry of the matrix A_{22} and $\operatorname{var}\tilde{v}_l$ is the lth diagonal entry of the matrix A_{44} (refer to Appendix B for the definitions of these matrices). A large standardized value of $\tilde{u}_k$ means that the kth individual has a relatively high odds of injury incidence. For an individual exhibiting a particularly high density of injury, a corresponding large standardized value of $\tilde{v}_l$ will be observed. Examination of these values assists in identifying individuals with high injury occurrence rate and/or high injury density once injured.

4.5. Goodness-of-fit and deviance residuals

To assess adequacy of the proposed model, we consider the scaled deviance test [19]. As indicated in Section 4.2, the log-likelihood can be separated into two independent components. Consequently, model assessment can also be undertaken with respect to the logistic and the truncated Poisson components. Specifically, the scaled deviance is defined by

$$D(y^*, \hat{\mu}^*) = -2[l(\hat{\mu}^*, \varphi \mid u, v) - l(y^*, \varphi \mid u, v)]$$

where φ is a scale parameter, y^* is the vector of responses with μ^* being its expected value, u and v are the random effects defined in the model, and $l(y^*, \varphi \mid u, v)$ defines the log-likelihood of y^* for given random effects u, v. Details are given in Appendix C. To detect systematic departure from the model, one may examine a plot of the deviance residuals, which are

defined as

$$\text{sign}(y_{ij}^* - \hat{\mu}_{ij}^*)\sqrt{\{-2[l(\hat{\mu}_{ij}^*, \varphi \mid u, v) - l(y_{ij}^*, \varphi \mid u, v)]\}}$$

4.6. Estimating mean injury incidence rate

In practical applications, it is often of interest to predict the mean injury incidence $E(Y_{ij}) = \mu_{ij}$ at specific covariate values. For the WRATS study, we are interested in both point and interval estimates of the mean injury incidence rate for male and female cleaners, before and after the intervention. Based on the ZIP random effects model, the mean injury incidence is estimated by

$$\hat{\mu}_{ij} = \frac{\hat{p}_{ij}\hat{\theta}_{ij}}{1 - \exp(-\hat{\theta}_{ij})}$$

With the time of exposure as offsets, $\hat{\theta}_{ij}$ and $\hat{p}_{ij}$ are defined as

$$\hat{\theta}_{ij} = t_{ij} \exp(\hat{\eta}_{ij})$$

$$\hat{p}_{ij} = \frac{t_{ij} \exp \hat{\xi}_{ij}}{1 + t_{ij} \exp \hat{\xi}_{ij}}$$

It should be noted that, in addition to fixed effects, random effects have been incorporated into the linear predictors. The approximate $100(1 - \alpha)$ per cent confidence interval for the mean injury incidence is $\hat{\mu}_{ij}(w_{ij}^*, x_{ij}^*) \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\{\text{var}\,\hat{\mu}_{ij}(w_{ij}^*, x_{ij}^*)\}}$, where Φ is the standard normal distribution function. Details on the derivation of $\text{var}(\hat{\mu}_{ij})$ are provided in Appendix D.

5. APPLICATION

The data comprised injury counts from a cohort of 137 cleaners who were present in pre- and post-WRATS intervention. Unlike the Poisson, the ZIP distribution does not indicate any lack of fit (details not presented here). To further control for confounding injury risk, covariates age and gender are included. For the ZIP regression model, covariates pre–post (binary indicator of WRATS intervention), gender and age are used in both the logistic and truncated Poisson submodels. In addition, the time of exposure is incorporated as an offset for both components. The time of exposure is expressed as per 365 working days. Results of the fitted model are given in Table II.

It is found that the WRATS intervention is significant in both logistic and truncated Poisson parts. This implies that the WRATS programme is effective in reducing the odds of injury incidence as well as reducing the incidence density for those non-zero observations. Moreover, gender is also significant in the logistic model, indicating that females have a higher odds of injury incidence. For variable selection, collinearity problem arises when interaction terms are incorporated, therefore only main effect models are considered. Results based on a deviance test for subset selection are given in Table III.

Table II. Parameter estimates of ZIP regression model with random effects.

Variables	Estimate (SE)	p-value
Logistic		
Intercept	−0.409 (0.693)	0.555
Pre–post	−1.263 (0.302)	0.000
Gender	−0.935 (0.403)	0.020
Age	0.008 (0.017)	0.638
σ_u^2	0.363 (0.367)	0.161
Truncated Poisson		
Intercept	−1.072 (0.851)	0.208
Pre–post	−1.186 (0.487)	0.015
Gender	0.148 (0.461)	0.748
Age	0.015 (0.021)	0.475
σ_v^2	0.372 (0.273)	0.086

Table III. Deviance tests for model selection.

Model	Logistic	Truncated Poisson
Pre–post, gender, age	268.71	61.66
Pre–post, gender	268.93	62.17
Pre–post, age	271.54	62.70
Gender, age	294.85	69.36
Pre–post	274.64	63.05
Gender	296.76	70.73
Age	299.78	70.34
Intercept only	301.09	72.11

Table IV. Parameter estimates of the final model.

Variables	Estimate (SE)	p-value	Odds ratio (95 per cent CI)
Logistic			
Intercept	−0.095(0.210)	0.651	
Pre–post	−1.221(0.288)	0.000	0.295 (0.168, 0.519)
Gender	−0.965(0.398)	0.015	0.381 (0.175, 0.831)
σ_u^2	0.363(0.367)	0.161	
Truncated Poisson			
Intercept	−0.463(0.162)	0.004	
Pre–post	−1.085(0.464)	0.019	0.338 (0.136, 0.839)
σ_v^2	0.351(0.263)	0.091	

Table IV presents the final model. The logistic part includes terms pre–post and gender, whereas the truncated Poisson part contains only the pre–post indicator. Therefore, the WRATS programme may be interpreted as effective in reducing both the injury incidence and the density once injured. For the logistic part, the variance of the random effect u is

Table V. Mean injury incidence (95 per cent CI) per 365 working days.

Group	Mean injury incidence (95 per cent CI)
(Pre, female)	0.642 (0.492, 0.791)
(Pre, male)	0.347 (0.155, 0.539)
(Post, female)	0.235 (0.149, 0.321)
(Post, male)	0.103 (0.028, 0.178)

estimated to be 0.363. A deviance goodness-of-fit test yields 268.93 on 255 degrees of freedom (p-value $= 0.263$). No particularly large deviance residual or standardized random effect is observed. For the truncated Poisson part, the variance of the random effect v is estimated to be 0.351. The deviance goodness-of-fit statistic has the value 63.05 on 78 degrees of freedom (p-value $= 0.890$). Again, there is no evidence of lack of fit and no outlying deviance residual is found. However, a large standardized random effect having value 2.94 is observed for one cleaner (case 111). This person sustained eight injuries in pre-WRATS but only a single injury in post-WRATS. Compared with the average injury count of 1.69 in the pre-period and 1.17 in the post-period for those ever injured, this cleaner apparently has experienced a much higher injury rate than others prior to the introduction of WRATS.

The mean injury incidence estimates and associated 95 per cent confidence intervals for the four covariate combinations are displayed in Table V. It is noted that while the mean injury incidence is higher for the female cleaners than their male counterparts, both groups are expected to sustain approximately one-third as many injuries post-intervention. Clearly, the intervention of the WRATS team has been of significant value.

6. DISCUSSION

A major feature of the WRATS study is that it was undertaken in a real workplace as opposed to a highly controlled but unrealistic laboratory environment. We have offered practical guidance to analyse longitudinal count data with extra zeros and provided statistical evidence to support the WRATS concept with respect to workplace safety. Taking into account the time of exposure and within-cleaner correlation, the conditional setting of the random effects ZIP regression model enables a convenient way to interpret the covariate effects directly through injury incidence and injury density in the respective logistic and truncated Poisson components. The methodology can also be applied to evaluate interventions targeted for populations at high risk of injury or harm from causes such as chemical, mechanical or biological hazards.

Given that ZIP models are special Poisson mixtures, a natural extension of embedding random effects within the framework of finite mixture GLM [20, 21] appears worthwhile. Formal tests remain to be developed, although empirical evidence [8, 10] indicates the ZIP model is often sufficient to explain the underlying population heterogeneity. As a final remark, each injury may be classified in four ways in accordance with WorkSafe Australia's *Type of Occurrence Classification System*, namely, nature, location (bodily), mechanism and agency. This suggests the generalization of the multivariate zero-inflated Poisson model [22] to the

random effects setting. Unfortunately, such refined classifications are not currently available due to historical reasons in the way the information was recorded.

APPENDIX A: SIMPLICATION OF THE FIRST AND SECOND DERIVATIVES

Further simplification leads to the following first and second derivatives:

$$\Im_{\alpha,u} = \begin{bmatrix} W' \\ Z_u' \end{bmatrix} \left(-\frac{\partial^2 l_1}{\partial \xi \partial \xi'} \right) [W \;\; Z_u] + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_u^{-2} I_n \end{bmatrix}$$

$$\Im_{\beta,v} = \begin{bmatrix} X' \\ Z_v' \end{bmatrix} \left(-\frac{\partial^2 l_1}{\partial \eta \partial \eta'} \right) [X \;\; Z_v] + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_v^{-2} I_m \end{bmatrix}$$

$$\frac{\partial l_\xi}{\partial \alpha} = W' \frac{\partial l_1}{\partial \xi}, \qquad \frac{\partial l_\xi}{\partial u} = Z_u' \frac{\partial l_1}{\partial \xi} - \sigma_u^{-2} u$$

$$\frac{\partial l_n}{\partial \beta} = X' \frac{\partial l_1}{\partial \eta}, \qquad \frac{\partial l_\eta}{\partial v} = Z_v' \frac{\partial l_1}{\partial \eta} - \sigma_v^{-2} v$$

where ξ, η are vectors of ξ_{ij} and η_{ij}, respectively; W, Z_u, X, Z_v are the design matrices for α, u, β and v, respectively. Now

$$l_1 = \sum_{y_{ij}=0} \log \left(\frac{1}{1 + t_{ij} \exp \xi_{ij}} \right) + \sum_{y_{ij}>0} \log \left(\frac{t_{ij} \exp \xi_{ij}}{1 + t_{ij} \exp \xi_{ij}} \right)$$

$$+ \sum_{y_{ij}>0} (y_{ij}(\log t_{ij} + \eta_{ij}) - t_{ij} \exp \eta_{ij} - \log(1 - e^{t_{ij}(-\exp \eta_{ij})}) - \log(y_{ij}!))$$

$$\frac{\partial l_1}{\partial \xi_{ij}} = \begin{cases} \dfrac{-t_{ij} \exp \xi_{ij}}{1 + t_{ij} \exp \xi_{ij}} & y_{ij} = 0 \\ \dfrac{1}{1 + t_{ij} \exp \xi_{ij}} & y_{ij} > 0 \end{cases}$$

$$= \frac{1}{1 + t_{ij} \exp \xi_{ij}} - D_{(y_{ij}=0)}$$

where $D_{(\,\cdot\,)}$ is an indicator function.

$$\frac{\partial l_1}{\partial \eta_{ij}} = y_{ij} + \frac{-t_{ij} e^{\eta_{ij}}}{1 - \exp(-t_{ij} e^{\eta_{ij}})} \quad \text{for } y_{ij} > 0$$

$$= y_{ij} + \frac{z_{ij}}{1 - \exp z_{ij}} \quad \text{where } z_{ij} = -t_{ij} e^{\eta_{ij}}$$

Let ξ, η, t, y, z be the corresponding vectors. We then have

$$\frac{\partial l_1}{\partial \xi} = \frac{1}{1 + t \exp \xi} - D_{(y=0)}$$

and

$$\frac{\partial l_1}{\partial \eta} = y + \frac{z}{1 - \exp z}$$

Hence

$$-\frac{\partial^2 l_1}{\partial \xi \partial \xi'} = \text{Diag}\left[\frac{t \exp \xi}{(1 + t \exp \xi)^2}\right]$$

$$-\frac{\partial^2 l_1}{\partial \eta \partial \eta'} = \text{Diag}\left[\frac{-z(1 - \exp z + z \exp z)}{(1 - \exp z)^2}\right]$$

APPENDIX B: ASYMPTOTIC VARIANCES AND VARIANCE COMPONENT ESTIMATES

Following the notation in Sections 4.2 and 4.3, suppose $\mathfrak{I}_{\alpha,u}$ is partitioned conformally to $\alpha|u$ as

$$\mathfrak{I}_{\alpha,u}^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

and $\mathfrak{I}_{\beta,v}$ is partitioned conformally to $\beta|v$ as

$$\mathfrak{I}_{\beta,v}^{-1} = \begin{bmatrix} A_{33} & A_{34} \\ A_{43} & A_{44} \end{bmatrix}$$

Asymptotic variances of the fixed effect parameter estimates are then given by $\text{var}\,\hat{\alpha} = A_{11}$ and $\text{var}\,\hat{\beta} = A_{33}$.

Variance components and their asymptotic variances are obtained by modifying equation (3.2) and the REML information matrix of McGilchrist [17]

$$\hat{\sigma}_u^2 = n^{-1}[\text{tr}\, A_{22} + \tilde{u}'\tilde{u}]$$

$$\hat{\sigma}_v^2 = m^{-1}[\text{tr}\, A_{44} + \tilde{v}'\tilde{v}]$$

$$\text{var}\,\hat{\sigma}_u^2 = 2[\sigma_u^{-4}(n - 2\sigma_u^{-2}\,\text{tr}\, A_{22}) + \sigma_u^{-8}\text{tr}(A_{22}^2)]^{-1}$$

$$\text{var}\,\hat{\sigma}_v^2 = 2[\sigma_v^{-4}(m - 2\sigma_v^{-2}\,\text{tr}\, A_{44}) + \sigma_v^{-8}\text{tr}(A_{44}^2)]^{-1}$$

APPENDIX C: COMPUTATION OF SCALED DEVIANCES

The scaled deviance follows an asymptotic $\chi^2_{n^*-t^*}$ distribution. For the logistic component, y^* is a vector of y^*_{ij}, where y^*_{ij} equals 0 or 1 according to the non-occurrence or occurrence of injury in period j for the ith individual with $n^* = n$ and $t^* = \mathrm{tr}(H_u^{-1}H_u^*)$, where $H_u = \Im_{\alpha,u}$ and $H_u^* = \Im_{\alpha,u} - \begin{bmatrix} 0 & 0 \\ 0 & \sigma_u^{-2} I_n \end{bmatrix}$. For the truncated Poisson component, y^* represents an $m \times l$ vector recording all non-zero injury counts with $n^* = m$ and $t^* = \mathrm{tr}(H_v^{-1}H_v^*)$, where $H_v = \Im_{\beta,v}$ and $H_v^* = \Im_{\beta,v} - \begin{bmatrix} 0 & 0 \\ 0 & \sigma_v^{-2} I_m \end{bmatrix}$.

APPENDIX D: VARIANCE OF MEAN INJURY INCIDENCE RATE

Let $\hat{\alpha}^* = \binom{\hat{\alpha}}{\hat{u}}, \hat{\beta}^* = \binom{\hat{\beta}}{\hat{v}}$ and $V = \mathrm{var}\binom{\hat{\alpha}^*}{\hat{\beta}^*} = \begin{pmatrix} \Im_{\alpha,u}^{-1} & 0 \\ 0 & \Im_{\beta,v}^{-1} \end{pmatrix}$. Since $\hat{\mu}_{ij}$ is a function of $\hat{\alpha}^*$ and $\hat{\beta}^*$, so $\mathrm{var}(\hat{\mu}_{ij})$ can be expressed as

$$\mathrm{var}(\hat{\mu}_{ij}) \simeq \begin{pmatrix} \dfrac{\partial \hat{\mu}'_{ij}}{\partial \alpha^*} & \dfrac{\partial \hat{\mu}'_{ij}}{\partial \beta^*} \end{pmatrix} \left(\mathrm{var} \begin{pmatrix} \hat{\alpha}^* \\ \hat{\beta}^* \end{pmatrix} \right) \begin{pmatrix} \dfrac{\partial \hat{\mu}_{ij}}{\partial \alpha^*} \\ \dfrac{\partial \hat{\mu}_{ij}}{\partial \beta^*} \end{pmatrix}$$

Now, $V = \mathrm{var}\binom{\hat{\alpha}^*}{\hat{\beta}^*}$

$$\frac{\partial \hat{\mu}_{ij}}{\partial \alpha^*} = \left(\frac{\hat{\theta}_{ij}}{1 - \exp(-\hat{\theta}_{ij})} \frac{\partial \hat{p}_{ij}}{\partial \xi_{ij}} \frac{\partial \xi_{ij}}{\partial \alpha^*} \right) = \left(\frac{\hat{\theta}_{ij}}{1 - \exp(-\hat{\theta}_{ij})} \hat{p}_{ij}(1 - \hat{p}_{ij}) w^*_{ij} \right)$$

$$\frac{\partial \hat{\mu}_{ij}}{\partial \beta^*} = \left(\hat{p}_{ij} \frac{\partial}{\partial \eta_{ij}} \left(\frac{\hat{\theta}_{ij}}{1 - \exp(-\hat{\theta}_{ij})} \right) \frac{\partial \eta_{ij}}{\partial \beta^*} \right)$$

$$= \left(\hat{p}_{ij} \frac{\hat{\theta}_{ij}}{1 - \exp(-\hat{\theta}_{ij})} \left(1 + \hat{\theta}_{ij} - \frac{\hat{\theta}_{ij}}{1 - \exp(-\hat{\theta}_{ij})} \right) x^*_{ij} \right)$$

where w^*_{ij} is the corresponding column vector of the transpose of the matrix $(W \; Z_u)$ and x^*_{ij} is the corresponding column vector of the transpose of the matrix $(X \; Z_v)$.

Hence, $\mathrm{var}(\hat{\mu}_{ij}) = \gamma'\gamma$, where

$$\gamma = V^{1/2} \begin{pmatrix} \dfrac{\partial \hat{\mu}_{ij}}{\partial \alpha^*} \\ \dfrac{\partial \hat{\mu}_{ij}}{\partial \beta^*} \end{pmatrix}$$

ACKNOWLEDGEMENTS

The authors thank Dr Phil Carrivick of the Department of Occupational Health, Sir Charles Gairdner Hospital, Western Australia, for discussion of the WRATS study. The computer program is stored in a Dyalog APL workspace available from the first author's web page: http://fbstaff.cityu.edu.hk/mskyau/. This research is supported in part by grants from the Research Grant Council of Hong Kong.

REFERENCES

1. WorkSafe Australia. *Occupational Health and Safety Performance*, *Overviews*, *Selected Industries. Issue No. 7 – Hospitals*, *Nursing Homes and Related Industries*. Australian Government Publishing Service: Canberra, 1995.
2. Snook SH, Campanelli RA, Hart JW. A study of three preventive approaches to low back pain injury. *Occupational Medicine* 1978; **20**:478.
3. Boucat R, Gun R, Ryan P. An evaluation of the risk identification checklist from the manual handling code of practice. *Journal of Occupational Health and Safety* 1994; **10**:205–211.
4. Bohr P, Evanoff B, Laurie D. Implementing participatory ergonomics teams among health care workers. *American Journal of Industrial Medicine* 1997; **32**:190–196.
5. Koda S, Ohara H. Preventive effects on low back pain and occupational injuries by providing the participatory occupational safety and health program. *Journal of Occupational Health* 1999; **41**:160–165.
6. McLachlan GJ. On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research* 1997; **6**:76–98.
7. Hinde J, Demétrio CGB. Overdispersion: models and estimation. *Computational Statistics & Data Analysis* 1998; **27**:151–170.
8. Böhning D, Dietz E, Schlattmann, P. Zero-inflated count models and their applications in public health and social science. In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, Rost J, Langeheine R (eds). Waxmann: Münster, 1997.
9. Johnson NL, Kotz S, Kemp AW. *Univariate Discrete Distributions*, 2nd edn. Wiley: New York, 1992.
10. Böhning D, Dietz E, Schlattmann P, Mendonça L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**:195–209.
11. Gupta PL, Gupta RC, Tripathi RC. Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis* 1996; **23**:207–218.
12. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**:1–14.
13. Van den Broek J. A score test for zero inflation in a Poisson distribution. *Biometrics* 1995; **51**:738–743.
14. Grogger JT, Carson RT. Models for truncated counts. *Journal of Applied Econometrics* 1991; **6**:225–238.
15. Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* 1996; **88**:297–308.
16. Schall R. Estimation in generalised linear models with random effects. *Biometrika* 1991; **78**:719–727.
17. Breslow NE, Clayton DG. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
18. McGilchrist CA. Estimation in generalised mixed models. *Journal of the Royal Statistical Society*, *Series B* 1994; **56**:61–69.
19. Lee Y, Nelder JA. Hierarchical generalized linear models. *Journal of the Royal Statistical Society*, *Series B* 1996; **58**:619–678.
20. Jansen RC. Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* 1993; **49**:227–231.
21. Thompson TJ, Smith PJ, Boyle JP. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Applied Statistics* 1998; **47**:393–404.
22. Li CS, Lu JC, Park J, Kim K, Brinkley PA, Peterson JP. Multivariate zero-inflated Poisson models and their applications. *Technometrics* 1999; **41**:29–38.

