Regression models for public health surveillance data: a simulation study

H Kim, D Kriebel

University of Massachusetts Lowell, Lowell, Massachusetts, USA

Correspondence to:
Hyun Kim, Department of Community and Preventive Medicine, Mount Sinai School of Medicine, One Gustave L Levy Place - Box 1057, New York, NY 10029, USA;
hyun.kim@mssm.edu

Accepted 5 April 2009
Published Online First 16 August 2009

ABSTRACT

Objectives: Poisson regression is now widely used in epidemiology, but researchers do not always evaluate the potential for bias in this method when the data are overdispersed. This study used simulated data to evaluate sources of overdispersion in public health surveillance data and compare alternative statistical models for analysing such data. If count data are overdispersed, Poisson regression will not correctly estimate the variance. A model called negative binomial 2 (NB2) can correct for overdispersion, and may be preferred for analysis of count data. This paper compared the performance of Poisson and NB2 regression with simulated overdispersed injury surveillance data.
Methods: Monte Carlo simulation was used to assess the utility of the NB2 regression model as an alternative to Poisson regression for data which had several different sources of overdispersion. Simulated injury surveillance datasets were created in which an important predictor variable was omitted, as well as with an incorrect offset (denominator). The simulations evaluated the ability of Poisson regression and NB2 to correctly estimate the true determinants of injury and their confidence intervals.
Results: The NB2 model was effective in reducing overdispersion, but it could not reduce bias in point estimates which resulted from omitting a covariate which was a confounder, nor could it reduce bias from using an incorrect offset. One advantage of NB2 over Poisson for overdispersed data was that the confidence interval for a covariate was considerably wider with the former, providing an indication that the Poisson model did not fit well.
Conclusion: When overdispersion is detected in a Poisson regression model, the NB2 model should be fit as an alternative. If there is no longer overdispersion, then the NB2 results may be preferred. However, it is important to remember that NB2 cannot correct for bias from omitted covariates or from using an incorrect offset.

Public health surveillance activities often collect data on adverse events in a count format and the outcomes are expressed as population rates, such as cases per 1000 population/year, or the number of sharps injuries per 100 full-time equivalent workers (FTEs)/year. In this type of surveillance, often called population-based surveillance,[1] the calculation of rates allows investigators to compare the experience of different groups, for example by geography, over different time periods or by occupation. Often rates are calculated using a person-time denominator, such as the population at risk in 1 year. But other types of denominators are possible and may make rates more useful for surveillance. For example, automobile crash data can be expressed as the rate per person-time or per the number of miles driven; the latter more closely reflects the exposure (to driving) and so accounts for differences in the driving habits of different classes of drivers.

What this paper adds

- Poisson regression is well-recognised in epidemiology as an appropriate statistical method for modelling count data as often occur in cohort studies calculating incidence rates.
- A serious challenge to the validity of the results of Poisson regression can arise if the count data are overdispersed, violating the Poisson equidispersion assumption.
- An alternative to Poisson regression, negative binomial regression, is now widely available in statistical software packages, but it has not been clear when the latter should be preferred to the former.
- Using simulation models, this paper showed that the negative binomial model was effective in reducing overdispersion in a wide range of scenarios, but it could not reduce bias in point estimates which resulted from omitting a covariate which was a confounder, nor could it reduce bias from using an incorrect offset, or denominator.
- The paper proposes steps to follow to evaluate overdispersion and choose between Poisson and negative binomial models.

Whatever the denominator, disease or injury rates are not usually assumed to have normal error distributions, and it is not generally appropriate to analyse them with ordinary least squares regression.[2] A more common approach is to think of the numerator of the rate – the number of events – as a count which follows the Poisson distribution.[2–5] Following this logic, one can model population-based surveillance data using Poisson regression, with the denominator of the rate included as an "offset". There is an important assumption in this approach, however; using the Poisson distribution to represent the error structure in the data assumes that the variance is equal to the mean – called *equidispersion*.[3–5] When this assumption is violated and the data exhibit *overdispersion*, inferences drawn from Poisson regression models will be incorrect and potentially quite misleading if the overdispersion is severe (the inverse problem – underdispersion – can occur, but it is much less common; it is not considered further in this paper).

In some kinds of surveillance data, overdispersion is common. For example, in Massachusetts, a legally mandated sharps injury surveillance program collects data on healthcare workers' sharps injuries from hospitals.[6] Annual reports are submitted without personal identifiers, and are used to generate information useful to guide prevention activities. The unit of analysis is an individual hospital, not a hospital employee. Large hospitals will typically have a larger number of injuries than small hospitals, and so when these data are viewed as counts at the hospital level, the distribution will have a "long right tail" meaning a few large hospitals have a large number of injuries while most hospitals have only a few. This kind of distribution will violate Poisson's equidispersion assumption. Poisson regression is now widely used in a variety of different types of epidemiological studies. But often the authors do not mention whether they checked for overdispersion in their data or whether they considered the sensitivity of their results as regards the choice of Poisson as their model.

The objectives of this paper were to evaluate sources of overdispersion (unobserved covariates and incorrect offset use) in population-based public health surveillance data, and compare regression models to identify methods which were suitable for analysing over-dispersed surveillance data. We used Monte Carlo simulation to develop datasets with varying degrees and sources of overdispersion and then compared the performance of different regression methods with these simulated data.

In this paper, we use the example of the Massachusetts Sharps Injury Surveillance System (MSISS) to motivate the methods that are evaluated.[6–8] Like many such surveillance programs, there is active data gathering for the events (needlesticks for example), while denominator and risk factor data are collected from existing administrative databases.

The Poisson regression model and overdispersion

The Poisson probability density function is:

$$f(y \mid \mu) = \frac{e^{-t\mu} t\mu^{y}}{y!} \qquad y = 0, 1, 2, \ldots$$

where the random variable y is a count response and non-negative integer, and the parameter μ is the mean or expected value of y. μ is also called the rate, t is usually the unit of observation time (person-time) but may be instead a unit of "exposure" like the number of miles driven in the case of automobile fatality rates.

Events which follow the Poisson distribution have a mean or expected value E[y] which is equal to the variance of those events, V[y]. This equidispersion assumption imposes an important restriction on the utility of Poisson regression. If the distribution of events in a surveillance dataset is over-dispersed, then the Poisson variance will be miss-specified; often the result will be an under-estimated or deflated standard error. The consequence will be overly optimistic statistical inferences in the sense of erroneously small p values and narrowed confidence intervals.[3]

From a statistical perspective, overdispersion is a violation of the assumption of independently identically distributed (IID) events.[3 5] A sequence of random variables is IID if each event follows the same distribution as the others and each is independent of the others. Violation of the IID assumption can occur when there is correlation (often but not necessarily positive) between responses, or excessive variation between response probabilities or counts. In the statistical literature, this is often thought to occur when an explanatory regression model omits important predictors; that is, there are unobserved covariates.[3–5]

Alternative models to correct for overdispersion

Over the past decade, alternative regression models to correct overdispersion have been introduced. One approach makes adjustments to the Poisson model to correct for overdispersion.[4] The other main strategy is to replace the Poisson with a negative binomial model. The negative binomial regression model can be derived from a mixture of the Poisson and gamma or exponential distributions,[3–5] allowing extra variance to be explicitly quantified within a variant of the Poisson model. The mass function of the negative binomial as a Poisson-gamma mixture is:

$$f(y) = \int_0^\infty Poisson(y \mid \mu) \cdot Gamma(\mu \mid r, (1-p)/p) d\mu$$

$$= \int_0^\infty \frac{e^{-\mu}\mu^{y}}{y!} \frac{\mu^{r-1} e^{-rp/(1-o)}}{\Gamma(r)} d\mu$$

where r is the gamma scale parameter, $p = r/(\mu+r)$ with $r>0$ and $0<p<1$, Γ is the gamma function, and the other parameters are same as in the Poisson probability density function (pdf) (equation 1).

The standard negative binomial model, called negative binomial 2 (NB2), has variance function $V = (1+\alpha\mu)\mu$. The overdispersion parameter α is an inverse gamma scale parameter (r).[3–5] When α is 0, the variance function V reduces to μ, as in the Poisson pdf, so that E[y] = V[y] = μ; which is the condition of equidispersion. Another negative binomial model, (NB1), has a slightly different variance function: $V = (1+\alpha)\mu$. Since NB2 is more widely used in the statistical literature, and often is the only variant implemented in statistical packages, it is the form that we consider here.

METHODS

Hypothesised sources of overdispersion in surveillance data

We hypothesised that overdispersion in surveillance data could be controlled by using an appropriate regression model. Potential sources of overdispersion examined in this investigation were: (1) unobserved covariates (the cause cited by most statistical texts); and (2) an incorrect offset or denominator for the grouped data. An extreme but common example of the latter is a failure to include any offset at all – to attempt to model counts without their denominator. Both these problems may result in violations of the IID assumption for count responses. Unobserved covariates probably occur frequently because public health agencies often lack data on important explanatory variables. When individual level counts following the Poisson distribution are grouped into categories, the groups become the units of analysis. One of the key properties of the Poisson distribution is the Countable Additivity theorem[5] which says that the sum of Poisson random variables is Poisson distributed. In the Poisson pdf, $y_{\text{individual}}$ becomes y_{group},

$$\sum_{1}^{i} y_{ij} = y_{j\sum_i} \sim \text{Poisson distribution}$$

where i represents individuals and j groups. Thus if individual level counts follow the Poisson distribution, then grouped counts also follow the Poisson distribution. However, when the y_i are not independent one from another, the y_j may not be

Poisson distributed.[5] One of the practical reasons for IID violation in the additivity theorem is that a mis-specified offset is used, so that the group level count data – even if they are generated by a Poisson process at the individual level – will not be Poisson distributed. Suppose for example that all individuals face the same risk of injury, and all are followed for the same period of time –for instance 1 year. Then at the individual level, *t* in equation 1 can be replaced by the value 1, and there is no need for an offset. At the group level, the unit of analysis becomes a group of individuals, and the offset becomes the number of individuals in each group. But now suppose that individuals in some groups are at risk for shorter periods than others. Then the offset of the number of people will not be correct. We hypothesised that this kind of error in offset might be an important reason for population-based surveillance data being overdispersed.

Simulation

We simulated datasets of sharps injuries (SI) among hospital workers following closely the patterns of data (but not the actual injury counts) observed in the MSISS during 2002–2003.[7] [8] MSISS covers all licensed hospitals in MA, and hospitals are required to report annually on SI to the Department of Public Health (MassDPH).[6] MSISS is a population-based passive surveillance system with a two stage reporting procedure. Each employee who experiences a SI is directed to report to the designated reporting system in their facility. Each facility collects these data and reports them annually in summary form to the MSISS. Along with each injury, there is information on the occupation and department of the case, as well as event details including the type of device. No individual identifiers are included. Also available to the MassDPH is a large amount of information on the reporting facility, including the type of hospital, location, number of employees, and number of licensed and occupied beds.

To mimic MSISS data, we created individual and group levels of risk factor information analogous to the individual and hospital data described above. We simulated datasets of SI in which there were two covariates. We imagined these to represent: (1) whether an injury was reported or not (variable REPORT), which could be thought of as either an individual level or group level factor but in either case is not measurable with surveillance data; and (2) whether the hospital was a teaching or a non-teaching facility (variable TEACH), which tends to affect the patient mix as well as many aspects of the safety climate of the hospital. The latter variable is publically accessible information, and we used the actual distribution of teaching and non-teaching hospitals in Massachusetts.

We generated datasets of 76 736 individual observations to mimic the total working population from the 64 acute care hospitals covered by MSISS. We assumed these individual observations to each represent a hospital FTE who either was or was not injured by a sharps device. With fixed values for the parameters determining the strength of the effects of REPORT and TEACH (see below), observations were determined to have had or not had a sharps injury according to a Poisson process:

$$Pois(\mu) = (\beta_R \times REPORT)+(\beta_T \times TEACH)+\varepsilon$$

where μ is SI response, $Pois(\,)$ is the Poisson pdf, and β_R and β_T are parameters for each covariate specifying their effects on the probability of injury, and REPORT and TEACH are an individual's (or his/her hospital's) values reflecting reporting behaviour and teaching/non-teaching status. We included ε, error, in the prediction model selected from a uniform distribution of pseudorandom numbers. For simplicity, we assumed that there were just two types of reporting behaviour – low and high likelihood to report an injury, and two types of hospitals, teaching and non-teaching. Each individual was therefore assigned a value of 1 or 2 for REPORT and each hospital was assigned a 1 (teaching) or 2 (non-teaching).

We generated 76 736 FTEs in each of 100 datasets. SI status was determined as described above. The 100 datasets varied in the strength of the two covariates β_R and β_T (table 1). The datasets were built with 10×10 sets of parameters,

$$\begin{pmatrix} \beta_{R,1} & \beta_{T,1} \\ \beta_{R,2} & \beta_{T,2} \\ \vdots & \vdots \\ \beta_{R,9} & \beta_{T,9} \\ \beta_{R,10} & \beta_{T,10} \end{pmatrix} = \begin{pmatrix} 0.01 & -1.5 \\ 0.02 & -1.6 \\ \vdots & \vdots \\ 0.09 & -2.3 \\ 0.1 & -2.4 \end{pmatrix}$$

The values of these parameters were chosen to reflect a realistic range of injury rates and their associations with these risk factors.

After the 76 736 FTEs and simulated SI were generated, they were collapsed into 64 hypothetical acute care hospitals. Each hospital was assigned a different number of FTEs, so that the true range of acute care hospital sizes in MA was created (from 59 to 6547 FTEs). By assuming that each observation was a full-time worker, the correct offset for hospital level analyses was defined to be the number of FTEs in the hospital. To investigate the effect of an incorrect offset, we used actual data on the number of licensed beds in each hospital. The range among the 64 hospitals was 19–912 beds. The number of beds was highly correlated with the number of FTEs (r = 0.89) (fig 1).

We created a correlation between hospital size and the covariates by assigning all FTEs in the 29 small hospitals (range of FTEs from 59 to 640) a low reporting behaviour (29 hospitals were coded 2 for REPORT meaning that their employees were half as likely to report their sharps injuries) and the larger 35 hospitals were assigned REPORT = 1 meaning less underreporting. As in the actual data, 14 hospitals were assigned to be teaching hospitals (TEACH = 2), and we assumed that this meant that they had more frequent use of sharps and a higher risk of injury.

Data analysis

Regression models were fit to the simulated datasets using generalised linear models. The variables REPORT and TEACH were used as independent variables, and the number of FTEs and the number of beds were used as offsets; models were also fit without any offset. Poisson and negative binomial (NB2) models were fit using Stata 9 (Stata, College Station, TX). To investigate the impacts and sources of overdispersion, six different models were compared:

- Model 1: a full-model including both covariates REPORT and TEACH, with correct offset (FTEs). This model assumed that there were no unobserved covariates and the offset was correct, so that there should be no overdispersion.
- Model 2: including both covariates but with incorrect offset (beds).
- Model 3: including both covariates but without offset.
- Model 4: omitting one covariate (REPORT), but with correct offset.

Table 1 Model configurations

Number of datasets	100									
Number of observations	76 736 FTEs									
Data collection period	12 months									
Underreporting (β_R)	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Teaching/non-teaching status (β_T)	−1.5	−1.6	−1.7	−1.8	−1.9	−2.0	−2.1	−2.2	−2.3	−2.4
Background risk (β_0)	0									
Error-term (ε)	−3 to 0 (uniformly distributed)									

FTE, full-time equivalent worker.

- Model 5: omitting one covariate (REPORT), and with incorrect offset.
- Model 6: omitting one covariate (REPORT), and with no offset.

Each model was fit to all 100 datasets generated by the simulations above.

The form of the Poisson regression equation for the full-model with the correct offset was:

$$\log(SI) = \log(FTE) + \beta_0 + \beta_R \times REPORT + \beta_T \times TEACH + \varepsilon,$$

which can also be written:

$$\log(SI/FTE) = \beta_0 + \beta_R \times REPORT + \beta_T \times TEACH + \varepsilon,$$

or again can be written:

$$SI/FTE = e^{(\beta_0 + \beta_R \times REPORT + \beta_T \times TEACH + \varepsilon)}$$

where SI is the number of simulated sharps injuries, FTE is the number of full-time equivalent workers and is used as the offset, SI/FTE is a Poisson rate, and the other terms have been previously defined. For each model, the following statistics were computed: $\hat{\beta}_R$ and $\hat{\beta}_T$, the estimated values of the two effects, and their standard errors, the deviance dispersion statistic (the value of the deviance divided by the degrees of freedom). The latter statistic indicates the degree of dispersion; values close to 1.0 indicate equidispersion, while values greater than 1.0 indicate overdispersion.

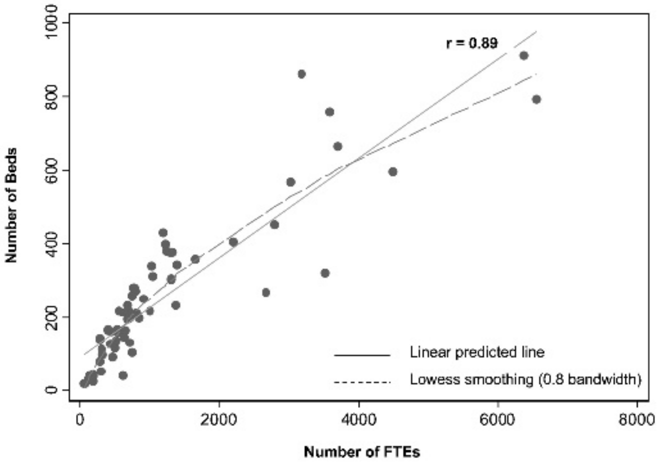

Figure 1 Comparing denominators (offsets) for sharps injury rates. Number of beds versus number of FTEs among 64 Massachusetts hospitals, 2002–2003. FTE, full-time equivalent worker.

RESULTS

The simulated data

Sharps injuries were generated first at the individual level. The mean number of SI across the 100 datasets was 3472 (range 1937–5786), which is close to the number of sharps injuries reported to MSISS from 64 acute care hospitals in 2003 (n = 3248). The values of the βs were chosen to obtain this result, so that the range of the data was realistic. The estimated β coefficients with individual level data differed slightly from their assigned true values because of random variation (table 2).

When individual level data were grouped into 64 hospitals, and the correct model was fit – including both covariates and using the correct offset – the estimated β coefficients were the same as those estimated at the individual level (compare scenario 1 with scenarios 2 and 3 in table 3). The coefficient for the second covariate, REPORT, was also correctly estimated using group level data, when the correct offset was used (data not shown). It made no difference whether Poisson or NB2 was chosen as the pdf, as one would expect for data which are not overdispersed. Deviance dispersion statistics were close to 1.0 for both group level models, indicating equidispersion. Grouping the injuries within the 64 hospitals, as would typically be done with such data, showed a typical right skewed distribution, consistent with equidispersed Poisson-distributed count data (fig 2).

Sources of overdispersion

The offset/denominator

Hospital-level regression models which included both covariates but used the number of beds as the offset rather than the number of FTEs, resulted in biased estimates of the effect of teaching versus non-teaching hospital (scenarios 4 and 5 in table 3). The estimated effect was about 0.60, while the true effect was much lower at 0.06. These coefficients can be translated into rate ratios comparing the sharps injury rates in teaching versus non-teaching hospitals (simulated data). The "true" rate ratio was RR = $e^{0.06}$ = 1.06, while the biased estimate resulting from using the wrong offset was RR = $e^{0.60}$ = 1.8. Also, the Poisson regression model (scenario 4) was overdispersed, as indicated by the deviance of 5.71 (fig 3). The NB2 version (scenario 5) was not overdispersed (deviance = 1.09), but the estimate of β for TEACH was just as biased as in the Poisson version. One advantage of the NB2 model is that it estimated a substantially larger standard error for the β, and hence a wider confidence interval. This reduced precision was appropriate, as the true effect was substantially smaller than the estimated effect (0.06 vs 0.60). Fitting models with no offset (scenarios 6 and 7) also resulted in even more biased estimates of the effect,

Table 2 Simulated number of sharps injuries, true values, estimated values and deviance dispersion statistics, computed by Poisson regression at the individual level

Dataset	Mean	SD	Min	Max
No of SIs*	3472*	1032	1937	5786
True β_R†	0.055	0.030	0.010	0.100
$\hat{\beta}_R$	0.058	0.044	-0.040	0.174
True β_T	-1.950	0.303	-2.400	-1.500
$\hat{\beta}_T$	-1.982	0.340	-2.811	-1.422
Deviance	0.278	0.060	0.182	0.393

*Number of sharps injuries (SIs) generated by simulation.
†True β_R is true value of β coefficient for REPORT; estimated $\hat{\beta}_R$ is estimated value of β coefficient for REPORT by simulation; true β_T is true value of β coefficient for TEACH; estimated $\hat{\beta}_T$ is estimated value of β coefficient for TEACH by simulation.

and again the NB2 model was not overdispersed while the Poisson model was.

Models with omitted covariate

Omitting the covariate for reporting behaviour introduced bias into the estimate of the other covariate, teaching/non-teaching status (scenarios 8 and 9). The omitted covariate was a confounder of the association between injury rate and teaching/non-teaching status, and it is well known that omitting a confounder from a regression model can introduce bias. The confounding occurred because as noted above, the simulated data included correlations between hospital size and both teaching/non-teaching status and hospital reporting behaviour. The latter was also an independent risk factor for injury rate, as a hospital with a lower reporting rate must necessarily have a lower number of (reported) injuries and hence a lower (reported) injury rate. Not surprisingly, changing the pdf from Poisson to NB2 did not decrease the bias caused by the omitted covariate (and in this particular instance actually increased the bias somewhat). An advantage of the NB2 model is illustrated in scenario 9: the confidence interval for the effect of teaching/non-teaching status is quite wide, and almost includes the true effect, while the Poisson model yielded a misleadingly tight confidence interval around the biased effect (scenario 8).

When models were constructed that included the two sources of overdispersion, an omitted covariate and an incorrect or absent offset, the bias increased still further, as did the deviance for the Poisson models (scenarios 10 through 13, fig 3). The NB2 models remained close to equidispersed.

Results comparing Poisson with NB2

As noted, all six alternative models were fitted to the 100 simulated datasets by both Poisson and NB2 (table 3). In the NB2 models, the overdispersion parameter α was calculated and used in determining the variance function in the Stata GLM procedure. Fitting NB2 instead of Poisson, substantially decreased overdispersion in all scenarios, except for the full model with correct offset (scenarios 2 and 3, table 3), in which the data were not overdispersed, meaning that $\alpha = 0$. In all other scenarios, α was significantly different from 0, and so the NB2 regression procedure used a variance function including a non-zero α. The resulting standard errors for β coefficients were increased and there was a concomitant reduction of deviance dispersion statistics. The wider confidence intervals for β estimated from NB2 rather than Poisson were a partial, but unsatisfactory compensation for the substantial biases that were introduced by either omitting a confounder or using the wrong offset.

DISCUSSION

Poisson regression is a widely-used method of assessing exposure–response associations in occupational and environmental epidemiology. We counted more than two dozen papers citing the method in this journal in the biennium 2007–2008. Many of these applications are classic cohort studies in which person-time data are available and are used as the denominator or offset for calculating rates. In this context, Poisson regression performs very well, either with grouped or individual-level data.[9] But when count data are drawn from cross-sectional

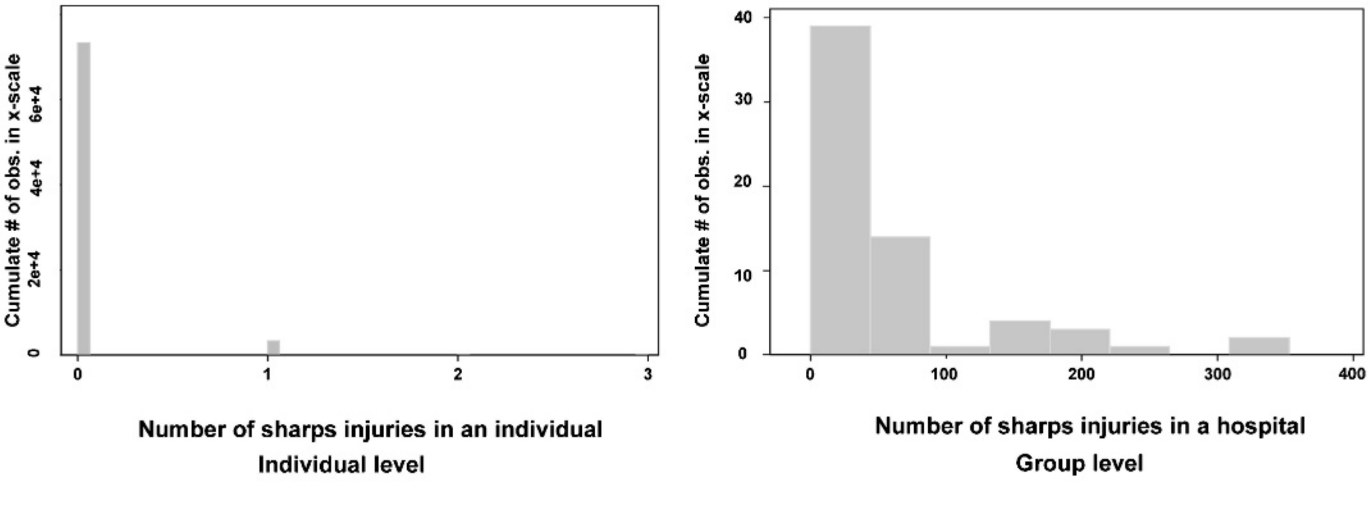

Figure 2 Histograms of the distributions of counts of sharps injuries for individuals and hospitals. obs., observed.

Table 3 Comparisons of results of alternative model scenarios

Scenario	Level	Covariates	Offset	Model	$\hat{\beta}_T$	SE	95% CI*	Bias†	Deviance
1	Individual	Including both covariates			0.06	0.04	−0.02 to 0.14	–	0.28
2	Group	Including both covariates	Correct	Poisson	0.06	0.04	−0.02 to 0.14	0	1.07
3				NB2	0.06	0.04	−0.02 to 0.14	0	1.07
4			Incorrect	Poisson	0.59	0.04	0.51 to 0.67	0.54	5.71
5				NB2	0.60	0.11	0.39 to 0.82	0.55	1.09
6			None	Poisson	1.05	0.04	0.97 to 1.13	0.99	17.99
7				NB2	1.03	0.18	0.68 to 1.38	0.97	1.11
8		Covariate REPORT omitted	Correct	Poisson	0.32	0.04	0.24 to 0.40	0.26	8.74
9				NB2	0.59	0.25	0.10 to 1.07	0.53	1.18
10			Incorrect	Poisson	0.89	0.04	0.81 to 0.97	0.83	13.82
11				NB2	1.13	0.26	0.62 to 1.65	1.08	1.17
12			None	Poisson	1.75	0.04	1.67 to 1.83	1.69	47.37
13				NB2	1.75	0.40	0.96 to 2.53	1.69	1.22

*From t statistic, ratio of β to standard error. Tests H_0: β = 0; †true minus observed β.
Individual level = 76 736 full-time equivalent workers; group level = 64 hospitals. SE, standard error.

surveys[10 11] or health surveillance databases,[12 13] denominator data are often missing or approximated from administrative data. In these situations, there is a risk of overdispersion with the potential for substantial errors in parameter estimation. Fortunately, negative binomial regression is now available in statistical packages, and offers considerable flexibility in fitting overdispersed datasets. We confirmed that the negative binomial model was effective in reducing overdispersion, but it could not reduce bias in point estimates which resulted from omitting a covariate which was a confounder, nor could it reduce bias from using an incorrect offset.

The problem of unmeasured confounders is well-known in epidemiology, and so it is not surprising that it affects analyses of surveillance data. Uncontrolled confounding did increase overdispersion as well as causing bias, and in these simulations, correcting for overdispersion in a model with uncontrolled confounding increased the width of confidence intervals for the covariates that were included in the model. A wide confidence interval may result from a poor model fit for a number of reasons, including omission of an important covariate. In these simulations, the coverage of the confidence interval for β from the NB2 models was closer (but still excluded) the true effect. Exactly how often the confidence interval will actually include the truth will depend on the magnitudes of effects of the omitted and included covariates and their correlations, but the principle of more conservative coverage from the NB2 model will generally hold.

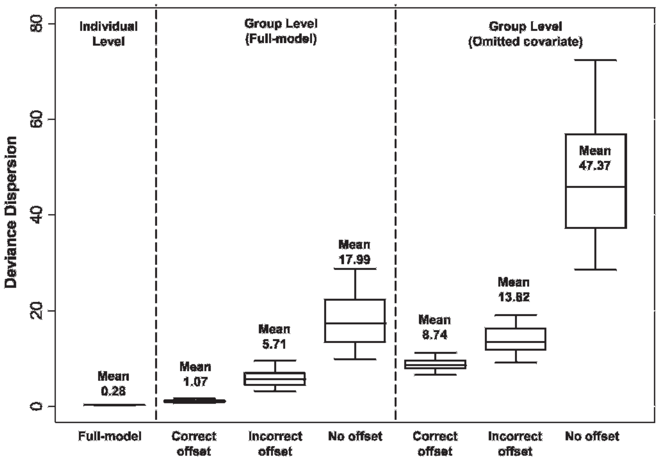

Figure 3 Mean Poisson deviance dispersion statistics for the 100 simulated datasets and six alternative models. Box and whisker plots indicate mean, inter-quartile range and minimum and maximum values.

The correct choice of denominators for calculating rates in surveillance data has been the subject of a number of investigations.[1 14–16] The denominator should be a measure of potential for exposure or injury, but there is no single figure that is an obvious choice. In this paper, we have shown that the incorrect denominator can actually introduce bias into regression models. It was surprising to find that using the number of hospital beds instead of the number of FTEs introduced substantial bias despite the fact that the two were highly correlated (r = 0.89, fig 1).

Analyses of surveillance data often ignore an offset all together, which led to serious bias. We believe that investigators may omit an offset from Poisson regression for at least two (mistaken) reasons. First, it is often assumed that risk may be constant among individuals. While this is a convenient assumption, it is probably often violated. In the case of hospital nurses at risk of sharps injuries, it is likely that exposure to sharps devices varies widely among nurses within the same hospital, which will be a strong determinant of risk. At the individual level, the correct offset is probably not the person or even the number of hours worked, but a measure of exposure.

The second mistake that may occur is to forget that analysis of group level data will require an offset, even if the individual level data do not. Suppose that risk actually is constant among all individuals, then an analysis at the individual level will show the data to be Poisson distributed, and no offset will be needed (or equivalently, the offset is 1.0). But when the data are aggregated into groups, an offset must be specified at the group level, and in this case, the correct offset for the grouped data will be the sum of persons in each group.

The negative binomial family of models and NB2 in particular, has been proposed as an alternative to Poisson regression which can correct for overdispersion.[3–5] The principal reason given for adjusting Poisson overdispersion is to minimise the risk of overly optimistic precision of effect estimates resulting from deflated standard errors and over-inflated t statistics.[3] These simulations did observe the expected effects when comparing results of Poisson and NB2 models. However, table 3 shows some mild overdispersion still remained in NB2 (scenarios 9, 11 and 13). Although not investigated here, a variety of adjustments and modifications of the negative binomial model have been proposed by Hilbe which may reduce overdispersion still further.[4]

One important limitation of this work is that the set of scenarios investigated was necessarily limited, and arbitrary choices were made about the magnitudes of the "true" effects which were simulated. Thus while the qualitative pattern of errors from mis-specified models should be broadly relevant, the magnitudes of the biases we observed may not be representative of real-world situations.

When analysing surveillance data that consist of counts of events, we recommend the following: fit the Poisson regression model first, and check for overdispersion using the deviance dispersion statistic. If there is overdispersion, then negative binomial models should be considered. If a negative binomial model still indicates overdispersion (deviance dispersion statistic $>>1.0$), model-based variance adjustment techniques should be considered within the negative binomial models.[4] It is important to remember though that neither omitted covariates nor incorrect offsets will be corrected by fitting NB2, and so one must emphasise the careful choice of covariates and denominators in the study design to minimise risk of bias.

Acknowledgements: The authors thank Dr Margaret Quinn and Dr Letitia Davis for their thoughtful advice.

Funding: This work was funded by the U.S. National Institute for Occupational Safety and Health (NIOSH) grant 1RO1 OH008229 to Project SHARRP, University of Massachusetts Lowell.

Competing interests: None.

Provenance and peer review: Not commissioned; externally peer reviewed.

REFERENCES

1. **Maizlish NA.** *Workplace health surveillance: an action-oriented approach*. New York: Oxford University Press, 2000.
2. **Hutchinson MK,** Holtman MC. Analysis of count data using Poisson regression. *Res Nurs Health* 2005;**28**:408–18.
3. **Cameron AC,** Trivedi PK. *Regression analysis of count data*. Econometric Society monographs, no. 30. Cambridge: Cambridge University Press, 1998.
4. **Hilbe JM.** *Negative binomial regression*. Cambridge: Cambridge University Press, 2007.
5. **Winkelmann RK.** *Econometric analysis of count data*. 3rd edn. Berlin: Springer, 2000.
6. **Office of Health and Human Services**. *Sharps Injury Surveillance and Prevention Project*. Available from http://www.mass.gov/?pageID=eohhs2terminal&L=5&L0=Home&L1=Consumer&L2=Community+Health+and+Safety&L3=Workplace+Health+and+Safety&L4=Needlesticks+and+Other+Sharps+Injuries&sid=Eeohhs2&b=terminalcontent&f=dph_occupational_health_c_needlesticks_sharps&csid=Eeohhs2 (accessed 27 December 2008).
7. **Massachusetts Department of Public Health**. *Sharps injuries among hospital workers in Massachusetts, 2002*. June 2004. Available from http://www.mass.gov/Eeohhs2/docs/dph/occupational_health/injuries_hospital.pdf (accessed 27 December 2008).
8. **Massachusetts Department of Public Health**. *Sharps injuries among hospital workers in Massachusetts, 2003*. January 2006. Available from http://www.mass.gov/Eeohhs2/docs/dph/occupational_health/injuries_hospital_2003.pdf (accessed 27 December 2008).
9. **Loomis D,** Richardson D, Elliott L. Poisson regression analysis of ungrouped data. *Occup Environ Med* 2005;**62**:325–9.
10. **Mirabelli M,** Zock J, Plana E, *et al*. Occupational risk factors for asthma among nurses and related healthcare professionals in an international study. *Occup Environ Med* 2007;**64**:474–9.
11. **Taimela S,** Laara E**,** Malmivaara A, *et al*. Self-reported health problems and sickness absence in different age groups predominantly engaged in physical work. *Occup Environ Med* 2007;**64**:739–46.
12. **Lauria L,** De Stavola BL. A district-based analysis of stillbirth and infant mortality rates in Italy: 1989–93. *Paediatr Perinat Epidemiol* 2003;**17**:22–32.
13. **Oviedo M,** Pilar Muñoz M, Domínguez A, *et al*. A statistical model to estimate the impact of a hepatitis A vaccination programme. *Vaccine* 2008;**26**:6157–64.
14. **Patel N,** Tignor GH. Device-specific sharps injury and usage rates: an analysis by hospital department. *Am J Infect Control* 1997;**25**:77–84.
15. **Jagger J.** Using denominators to calculate percutaneous injury rates. *Adv Expos Prev* 2002;**6**:7–8.
16. **Bena JF,** Bailer AJ, Loomis D. Effects of data limitations when modeling fatal occupational injury rates. *Am J Ind Med* 2004;**46**:271–83.