Epidemiol. Infect. (2008), **136**, 483–491. © 2007 Cambridge University Press
doi:10.1017/S0950268807008904 Printed in the United Kingdom

Epidemiological characteristics of reported sporadic and outbreak cases of *E. coli* O157 in people from Alberta, Canada (2000–2002): methodological challenges of comparing clustered to unclustered data

D. L. PEARL[1]*, M. LOUIE[2], L. CHUI[3], K. DORÉ[4], K. M. GRIMSRUD[5], S. W. MARTIN[1], P. MICHEL[6], L. W. SVENSON[5] AND S. A. McEWEN[1]

[1] *Department of Population Medicine, University of Guelph, Guelph, Ontario, Canada*
[2] *Provincial Laboratory for Public Health (Microbiology), Calgary, Alberta, Canada*
[3] *Provincial Laboratory for Public Health (Microbiology), Edmonton, Alberta, Canada*
[4] *Foodborne, Waterborne and Zoonotic Infections Division, Public Health Agency of Canada, Guelph, Ontario, Canada*
[5] *Alberta Health and Wellness, Edmonton, Alberta, Canada*
[6] *Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Saint-Hyacinthe, Québec, Canada*

(*Accepted 4 May 2007; first published online 13 June 2007*)

SUMMARY

Using multivariable models, we compared whether there were significant differences between reported outbreak and sporadic cases in terms of their sex, age, and mode and site of disease transmission. We also determined the potential role of administrative, temporal, and spatial factors within these models. We compared a variety of approaches to account for clustering of cases in outbreaks including weighted logistic regression, random effects models, general estimating equations, robust variance estimates, and the random selection of one case from each outbreak. Age and mode of transmission were the only epidemiologically and statistically significant covariates in our final models using the above approaches. Weighing observations in a logistic regression model by the inverse of their outbreak size appeared to be a relatively robust and valid means for modelling these data. Some analytical techniques, designed to account for clustering, had difficulty converging or producing realistic measures of association.

INTRODUCTION

Humans infected with *E. coli* O157 may show a range of clinical outcomes including asymptomatic shedding, gastroenteritis, haemorrhagic colitis, and/or haemolytic uraemic syndrome [1]. Infection with *E. coli* O157 has been associated with exposure to contaminated food [2, 3], contaminated drinking and recreational water [4, 5], shedding humans and animals [6, 7], and contaminated environments [8]. The young and elderly are often considered to be at the greatest risk of infection and/or complications following infection with *E. coli* O157 [1, 9–11]. Due to the potential severity of clinical symptoms, the level of underreporting of clinical cases is lower than disease associated with many other enteric pathogens [12, 13]. Current knowledge concerning risk factors associated with *E. coli* O157 comes generally from the analysis of databases on outbreaks and sporadic cases kept by public health agencies [14–16].

In a previous study in Alberta, Canada, we found variation in the spatial distribution of reported outbreak and sporadic cases [17]. We also demonstrated that the location of areas with significantly higher rates of *E. coli* O157, after correcting for demographic factors, was impacted by the choice to use only

* Author for correspondence: Dr D. L. Pearl, Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.
(Email: dpearl@uoguelph.ca)

sporadic cases or both sporadic and outbreak cases in our statistical analyses. Subsequently, we questioned whether the mode and site of transmission and the characteristics of individuals that became infected with *E. coli* O157 were different depending on whether they were identified as a sporadic or outbreak case (i.e. two or more cases related to a common exposure). Understanding these differences may have important implications for which groups and activities should be targeted for disease prevention programmes or increased surveillance. Studies on secondary transmission of *E. coli* O157 and a number of outbreaks dominated by children suggest young children may be at increased risk of being involved in outbreaks [6, 18].

Comparing outbreak and sporadic cases may pose some analytical challenges. Cases associated with a particular outbreak are not independent. Ignoring this lack of independence can result in an underestimation of variability and increase the probability of making a type I error (i.e. falsely rejecting the null hypothesis) [19–21]. It can also result in incorrect estimates of measures of association. A variety of techniques including robust variance estimates, general estimating equations (GEE), and random effects models have been created to deal with clustered data [19–21]. However, in our study, only cases associated with outbreaks had a true hierarchical structure. Consequently, methods, such as GEE or random effects models, that correct both measures of association and measures of variance by accounting for the correlation among cases, may fail to converge or provide unreliable estimates if faced with few replications at a hierarchical level. Other possible techniques for controlling for clustering may include the random selection of one case from each outbreak to remove the clustering effect by design rather than through a statistical technique. Analytical weights could also be used to down-weight the impact of multiple cases from a single outbreak; similar approaches have been developed for dealing with complex survey designs where the probability of being sampled may vary among cases [21].

Using a database of reported cases of *E. coli* O157 from Alberta, Canada in 2000–2002, we determined whether there were significant differences between sporadic and outbreak cases in terms of age, sex, the mode of transmission (e.g. food), and the site of transmission (e.g. home). We also determined whether the following temporal, spatial, and administrative factors were significant predictors of whether a case was more likely to be part of an outbreak: a region in the province where there was a higher proportion of outbreak cases than sporadic cases (outbreak cluster) [17]; a yearly period from May to October when reported cases cluster [17]; year of study; and whether a case had been hospitalized. We also tested the significance of interactions between age, sex, and mode and site of transmission. We created multivariable logistic models using a variety of methods to correct for clustering for comparative purposes.

METHODS

Data

We obtained a list of all reported cases of *E. coli* O157 infection in Alberta during the 2000–2002 period. The data were obtained from Notifiable Disease Reports (NDR) stored electronically in the Communicable Disease Reporting System (CDRS) maintained by the Disease Control and Prevention Branch of Alberta Health and Wellness using methods designed to preserve patient anonymity. The protocol for this research was approved by the University of Guelph Research Ethics Board.

In total 875 cases were recorded during this period, but six cases were excluded from subsequent analyses because of insufficient demographic and/or spatial data [17]. From the NDR we obtained/calculated the following information for each case: a unique identifier (NDR number), age at onset of clinical symptoms, year case occurred, sex, mode of transmission, site of transmission, case hospitalization (yes/no), a unique identifier used for community outbreaks (Exposure Indicator Number), cases identified through an epidemiological link that did not require definitive laboratory results (EPI-linked), the NDR numbers that connected EPI-linked cases, and a unique identifier created to identify cases sharing a common address. Mode of transmission included the following categories: food and water, animal contact, person-to-person, other, and unknown. Site of transmission included the following categories: home, food services establishment, high-contact housing (nursing home, daycare, campsite), travel, other, unknown. The categories for mode and site of transmission were slightly modified from the categories used in the NDR to avoid small cell sizes in subsequent statistical analyses. The ‘other’ category in the mode and site of transmission variables were known/suspected by the

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

public health worker, but did not fall into the remaining categories. In the electronic database, only 'other' was recorded, but we added three observations, believed to be attributed to the pathogen being aerosolized, to the 'other' category for mode of transmission, and 20 cases believed to have occurred at the workplace to the 'other' category in site of transmission. Cases were classified as outbreak cases if they shared a common Exposure Indicator Number, were Epi-linked, and/or shared a common address. Based on the above definition, cases associated with an outbreak were given a unique outbreak number. In total, there were 69 outbreaks involving 185 cases. The largest outbreak recorded in the database included 10 cases. However, 53% of all outbreak cases were in outbreaks involving two individuals.

Two variables in this study were based on the results of previous analyses of these data using SaTScan version 3.1.2 [22]. We created variables to identify a region in the province with an increased proportion of outbreak cases relative to sporadic cases based on a Bernoulli model (outbreak cluster), and a period of time variable when the rate of cases was increased based on a Poisson model (seasonal cluster) [17, 22]. Cases that were classified as being part of the outbreak cluster were located in a region with an approximate latitude and longitude of 50° N and 112° W, and a radius of 78 km. Cases that were classified as being part of a seasonal cluster fell between 1 May and 31 October in 2000–2002.

Statistical analyses

All of the multivariable statistical models created for this study involved a dichotomous outcome (i.e. outbreak/sporadic). The assumed distribution of the response variable was binary, and a logit link was used between the response variable and its linear predictor. All tests performed were two-tailed tests with a statistical significance level of 5%. In addition to descriptive statistics, univariable statistics were also performed for each variable with a *t* test, adjusted for unequal variances, being performed for age and χ^2 tests being performed for categorical variables.

Prior to beginning the model-building procedure, we determined whether there was a strong correlation among any of the predictor variables. If the correlation between two variables was $\geqslant 0{\cdot}8$, based on a Pearson's or Spearman's rank correlation coefficient, only one of the variables would be used in subsequent analyses. In addition, we tested the assumption of linearity for age using three techniques: including a square term in the model; categorizing the continuous predictor to determine if the coefficients increased in a relatively uniform fashion; and plotting the log odds of the outcome against the mean of 10 categories of age divided into 10 percentiles. If any of these methods revealed that the assumption of linearity was violated, the appropriate transformation was performed or if significant, a square term was added to the model.

The initial model-building procedure began with an ordinary logistic regression model (i.e. a logistic regression model that did not account for clustering in the data). Interaction terms and square terms were added individually to the full model (i.e. a model with all the main effects) and were removed if they were not statistically significant. Main effects were removed one at a time and kept in the final model if they were statistically significant based on a likelihood ratio test or acted as a confounding variable. A confounding variable was defined as a variable that caused at least a 20% change to the coefficient of a statistically significant variable on a log odds scale when it was removed from the model. All main effects that had been removed were re-introduced to the final model one at a time to re-assess their significance and potential confounding effect. When re-introducing interaction effects to a tentative final model, all necessary base terms were also included. The Hosmer–Lemeshow goodness-of-fit test was performed to determine that the binary model was appropriate for the data. A likelihood ratio test was used to assess the significance of removed variables where the estimates were calculated using maximum-likelihood procedures. Where a quasi-likelihood method was used, significance was assessed using Wald tests.

While our initial modelling procedures involved an ordinary logistic regression model, we looked at a number of procedures to correct for the impact of clustering [19, 21]. We employed the following two techniques that adjusted the variance estimates but not the coefficient estimates themselves: robust variance estimates based on Huber–White/sandwich variance estimates that are less sensitive to the assumption of independence; and robust variance estimates with a 'cluster' option that specifies that observations are independent across groups (i.e. cluster) but not necessarily within groups [21]. The following models adjusted both the coefficient estimates and the variances: a weighted logistic regression model where a weight was applied to each observation that was the inverse of the number of cases within an outbreak or

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

one if the case was sporadic; GEE with an exchangeable correlation structure; and random intercept models using various algorithms [21]. Random intercept models were performed/attempted using: various combinations of predictive quasi-likelihood (PQL) or marginal quasi-likelihood (MQL), and first- or second-order derivatives of the Taylor series expansion for linearization prior to the use of re-weighted iterative generalized least squares (RIGLS); the Monte Carlo Markov Chain (MCMC) method; and generalized linear latent and mixed models (GLLAMM) [20, 23]. In addition, we controlled for clustering prior to building a logistic model by randomly selecting one case from each outbreak. This procedure was performed five times to determine whether the coefficients calculated for each variable were relatively stable among iterations.

Residual analyses were performed for the ordinary logistic regression model and the random effects model produced using RIGLS, PQL, and first-order derivatives of the Taylor series expansion. Residuals were inspected for extreme standardized residuals, high leverage, and high influence. All statistical analyses were performed using Intercooled Stata 8.0 for Windows (Stata Corp., College Station, TX, USA) except for the random effects models involving quasi-likelihood or MCMC methods which were performed using MLwiN 2.02 (Institute of Education, London, UK). The weights for the weighted logistic regression were applied using an 'iweight' option. The weights were applied to the likelihood function in a manner similar to those used to adjust for different probabilities of being sampled [24].

RESULTS

Based on univariable statistics the following variables had a statistically significant association with the dependent variable (i.e. outbreak/sporadic): age, outbreak cluster, case year, mode and site of transmission (Table 1). The correlations among all the independent variables used in our study never exceeded 0·8. Age was log transformed (natural log) since it did not have a linear relationship with the outcome based on a log odds plot.

Within the ordinary logistic regression model the following variables were not statistically significant based on a likelihood ratio test ($\chi^2=11{\cdot}81$, D.F. $=11$, $P=0{\cdot}38$) and were removed from the full model: sex, hospitalization, case year, temporal cluster, and site of transmission. These terms did not have a confounding effect on the remaining significant variables. A square term for the log of age in years ($\chi^2=3{\cdot}13$, D.F. $=1$, $P=0{\cdot}077$), and interaction terms among age and sex ($\chi^2=0{\cdot}02$; D.F. $=1$; $P=0{\cdot}90$), age and mode of transmission ($\chi^2=3{\cdot}54$, D.F. $=4$, $P=0{\cdot}47$), sex and mode of transmission ($\chi^2=3{\cdot}23$, D.F. $=4$, $P=0{\cdot}52$), age and site of transmission ($\chi^2=4{\cdot}13$, D.F. $=5$, $P=0{\cdot}53$), and sex and site of transmission ($\chi^2=6{\cdot}47$, D.F. $=5$, $P=0{\cdot}26$) were also statistically non-significant when added to the full ordinary logistic regression model. The model-building process for the various logistic models designed to control for clustering resulted in models with the same fixed effects.

The final main effects models included the following variables: log of age in years, outbreak cluster, and mode of transmission (Table 2). The Hosmer–Lemeshow goodness-of-fit test for the ordinary logistic regression model was non-significant (Hosmer–Lemeshow $\chi^2=5{\cdot}20$, D.F. $=10$, $P=0{\cdot}74$) indicating that the binary model was appropriate. The direction of odds ratios for all the variables was the same among all the models (Table 2). The confidence intervals were generally larger for all models using methods designed to account for clustering (Table 2). However, applying the robust option to relax the assumption of independence among observations often resulted in narrower confidence intervals (Table 2) than those produced by our ordinary logistic regression model. Among methods designed to adjust the odds ratios and variance estimates, the random effects model employing PQL and the weighted regression model had almost identical odds ratios and 95% confidence intervals (Table 2). The odds ratios estimated using MQL were closer to the null among the statistically significant coefficients. The direction of odds ratios for significant variables was similar among the five models (Random 1–5) where a single case was randomly selected from each outbreak, although there was some variation in the point estimates and confidence intervals (Table 2). The model using GEE with an exchangeable correlation structure failed to converge. Estimates obtained using GLLAMM and MCMC methods when converted to an odds scale were several orders of magnitude greater or smaller, depending on the direction of the association, from those obtained using the other methods.

Increasing age appeared to have a statistically significant sparing effect, while cases found within the Bernoulli cluster were at increased risk of being outbreak cases (Table 2). Using food and water as the referent category, individuals believed to have

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

Table 1. *Univariable statistics comparing reported outbreak and sporadic cases from Alberta, Canada in 2000–2002*

Variable	Sporadic	Outbreak	Test statistic	*P* value
Age (years)	28·68 (23·58)	14·39 (17·64)	9·05	<0·001
Sex				
Male	291 (42·54)	79 (42·70)		
Female	393 (57·46)	106 (57·30)	0·0015	0·97
Hospitalized				
Yes	204 (29·82)	44 (23·78)		
No	435 (63·60)	130 (70·27)		
Unknown	45 (6·58)	11 (5·95)	2·95	0·23
Outbreak cluster				
Yes	66 (9·65)	50 (27·03)		
No	618 (90·35)	135 (72·97)	38·02	<0·001
Case year				
2000	258 (37·72)	66 (35·68)		
2001	236 (34·50)	49 (26·49)		
2002	190 (27·78)	70 (37·84)	7·94	0·019
Seasonal cluster				
Yes	531 (77·63)	155 (83·78)		
No	153 (22·37)	30 (16·22)	3·32	0·069
Mode of transmission				
Food and water	370 (54·09)	62 (33·51)		
Animal contact	52 (7·60)	10 (5·41)		
Person-to-person	31 (4·53)	61 (32·97)		
Other	12 (1·75)	6 (3·24)		
Unknown	219 (32·02)	46 (24·86)	128·65	<0·001
Site of transmission				
Home	286 (41·81)	103 (55·68)		
Food services	84 (12·28)	9 (4·86)		
High contact	17 (2·49)	20 (10·81)		
Other	51 (7·46)	11 (5·95)		
Travel	50 (7·31)	9 (4·86)		
Unknown	196 (28·65)	33 (17·84)	42·78	<0·001

For categorical variables, the number of cases and their relative proportions (in parentheses) are indicated. In the case of continuous variables (i.e. age), the mean and standard deviation (in parentheses) are indicated. *t* tests and χ^2 statistics were used for continuous and categorical variables, respectively.

obtained their infection from another person were at increased risk of being in an outbreak (Table 2). We also found significant differences in animal ($\chi^2=27{\cdot}10$, D.F. = 1, $P<0{\cdot}001$), other ($\chi^2=9{\cdot}24$, D.F. = 1, $P=0{\cdot}002$), and unknown ($\chi^2=43{\cdot}23$, D.F. = 1, $P<0{\cdot}001$) modes of transmission relative to person-to-person spread based on Wald tests applied to the ordinary logistic regression model. However, there was no significant difference between person-to-person transmission and the 'other' category for transmission when the following models/methods were employed to adjust for clustering: weighted regression ($\chi^2=2{\cdot}53$, D.F. = 1, $P=0{\cdot}11$); random effects model with PQL ($\chi^2=2{\cdot}22$, D.F. = 1, $P=0{\cdot}14$); random effects model with MQL ($\chi^2=0{\cdot}68$, D.F. = 1, $P=0{\cdot}41$); randomly selecting a case from each outbreak ($\chi^2=0{\cdot}52$–$2{\cdot}46$, D.F. = 1, $P=0{\cdot}12$–$0{\cdot}47$). There was no statistically significant relationship among the other modes of transmission. The residual analyses based on the ordinary and random effects logistic regression models did not identify any observations that were highly unusual or that had a large impact on the model when removed.

DISCUSSION

Age, mode of transmission, and a variable marking the spatial location of a cluster of outbreak cases (i.e.

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

Table 2. *The odds ratios and 95% confidence intervals (in parentheses) for all variables in our final multivariable models, using various methods to deal with clustering, comparing reported outbreak and sporadic cases from Alberta, Canada in 2000–2002*

Models/variables	Age in log years	Outbreak cluster	Mode, animal	Mode, other	Mode, person	Mode, unknown	Random variance
Logistic	0·65 (0·56–0·76)	2·83 (1·74–4·59)	0·73 (0·34–1·58)	1·15 (0·38–3·48)	6·81 (3·96–11·69)	1·02 (0·65–1·58)	n.a.
Robust	0·65 (0·56–0·76)	2·83 (1·77–4·51)	0·73 (0·35–1·55)	1·15 (0·42–3·12)	6·81 (3·95–11·72)	1·02 (0·65–1·59)	n.a.
Robust cluster	0·65 (0·56–0·76)	2·83 (1·43–5·62)	0·73 (0·30–1·76)	1·15 (0·39–3·39)	6·81 (3·44–13·48)	1·02 (0·57–1·81)	n.a.
Weighted	0·67 (0·55–0·83)	2·51 (1·27–4·97)	0·76 (0·25–2·33)	1·42 (0·32–6·28)	4·91 (2·31–10·44)	1·04 (0·55–1·97)	n.a.
PQL, 1st order, RIGLS	0·67 (0·54–0·82)	2·90 (1·36–6·17)	0·74 (0·25–2·18)	1·41 (0·29–6·85)	4·87 (2·18–10·88)	1·00 (0·54–1·85)	2·96 (1·73–4·18)
MQL, 1st order, RIGLS	0·74 (0·61–0·89)	2·35 (1·14–4·85)	0·81 (0·29–2·26)	1·33 (0·30–5·87)	2·50 (1·17–5·34)	1·02 (0·57–1·84)	3·92 (2·70–5·14)
Random 1	0·62 (0·50–0·76)	2·57 (1·27–5·19)	0·50 (0·14–1·78)	1·37 (0·33–5·72)	4·45 (2·12–9·32)	0·80 (0·41–1·56)	n.a.
Random 2	0·61 (0·49–0·75)	2·69 (1·36–5·33)	0·62 (0·20–1·92)	0·82 (0·16–4·21)	3·05 (1·40–6·65)	0·83 (0·44–1·56)	n.a.
Random 3	0·66 (0·54–0·81)	2·67 (1·35–5·28)	0·64 (0·21–1·96)	0·89 (0·17–4·53)	3·42 (1·58–7·42)	0·84 (0·45–1·58)	n.a.
Random 4	0·71 (0·58–0·88)	2·47 (1·26–4·85)	0·90 (0·32–2·52)	2·16 (0·60–7·76)	3·56 (1·60–7·90)	0·98 (0·52–1·84)	n.a.
Random 5	0·69 (0·56–0·85)	2·48 (1·25–4·90)	0·75 (0·24–2·31)	1·77 (0·44–7·17)	5·21 (2·47–10·98)	0·99 (0·52–1·89)	n.a.

n.a., Not applicable; Logistic, an ordinary logistic regression; Robust, robust variance estimates; Robust cluster, robust variance estimates accounting for the hierarchical structure; Weighted, weighted regression; PQL, predictive quasi-likelihood; RIGLS, re-weighted iterative generalized least squares; MQL, marginal quasi-likelihood; 1st order, first order derivative of the Taylor series expansion; Random (1–5), five iterations of a logistic regression where one case is randomly selected from each outbreak; Random variance, variance component of the random intercept.

outbreak cluster) were the only significant variables in our final models. The proportion of female and summer cases appeared to be higher than male and non-summer cases respectively during the 2000–2002 period (Table 1), but there appears to be no strong association between these variables and whether a case is sporadic or part of an outbreak. This suggests that the mechanisms behind determining whether a case is independent or part of an outbreak are not simply a reflection of risk factors associated with disease.

On a log scale, increasing age appeared to have a sparing effect on the risk of a case being linked to another case in the CDRS. This could be explained by the increased tendency of young children to sample their environment orally and practice poorer personal hygiene. There are many examples of outbreaks of *E. coli* O157 among young children associated with secondary transmission among children or exposure to shedding animals in farms or petting zoos [6, 7, 25]. A potential alternative hypothesis could be that the severity of disease in young children increases the effort or ability of health workers to make epidemiological links among cases [26]. However, cases that were hospitalized were not more likely to be linked to an outbreak, and there was no evidence to suggest that the risk of being linked to an outbreak increased again with advanced age when the impact of infection with *E. coli* O157 can be more severe [27].

Within the mode of transmission, only person-to-person transmission was significantly different from other categories. Depending on the model, a case was almost five times more likely to be within an identified outbreak if the public health worker believed that the mode of transmission was person-to-person compared to cases where the mode of transmission was related to food or water. Based on our definition of an outbreak (i.e. two or more cases linked epidemiologically), this finding was not surprising. In fact, it may be more important to note that nearly a third of cases where the mode of transmission was believed to be person-to-person remained unlinked to another case in the CDRS (Table 1). This may suggest that many outbreaks or cases associated with an outbreak remain unlinked or undetected. However, it is important to recognize that the sensitivity and specificity of questions concerning the mode and site of transmission of disease are likely to be different between truly sporadic and outbreak cases (i.e. differential misclassification bias). By virtue of being independent, the investigation of sporadic cases cannot be

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

supported with statistical measures used to investigate outbreaks [2, 6, 28]. As a result, we may be less certain about these classifications for sporadic cases.

Clustering or the lack of independence among observations is an important issue for obtaining appropriate measures of association and variance. In our study, clustering was an issue in terms of space and the grouping of cases into outbreaks. The inclusion of a variable to control for cases located within a spatial cluster of outbreak cases, identified in an earlier study [17], removed the possibility that our results were simply being driven by the clustering of 27% of our outbreak cases into a relatively small geographical area. A variety of more complex techniques are available to adjust for spatial autocorrelation [29], but we were able to create a simpler model using a fixed effect that was well delineated from our previous study [17]. On the other hand, modelling our outbreaks as fixed effects would have resulted in the introduction of a large number of variables into our model and the parameter estimates would have become unstable with relatively few observations per group [21]. In addition, we were faced with the unusual circumstance where sporadic cases, unlike outbreak cases, did not have a true hierarchical structure. Consequently, we wanted to compare a number of options for handling clustered data.

To create the most extreme scenarios, we created an ordinary logistic regression model that did not control for clustering, and five models where each time we only selected one case per outbreak (Table 2). Both types of models gave us reasonable expectations concerning the direction and relative size of the measures of association. Consequently, we were able to reject models generated by MCMC methods and GLLAMM that had strikingly different measures of association. However, these two models had measures of association that were effectively reaching positive and negative infinity and would not have been accepted as biologically plausible. In addition, the GEE model and the random effects model using the second-order derivative of the Taylor series expansion failed to converge. The poor performance of some of the models was expected by virtue of the structure of the data. Random effects and GEE models are often used to adjust for a correlation structure among grouped observations. Typically, all the data have a hierarchical structure although the number of observations within a higher level of the hierarchy may vary. However, more than half of our data (i.e. sporadic cases) had no true hierarchical structure, although they were each assigned a grouping variable to perform these procedures. In addition, the correlation structure among observations within an outbreak was defined by our dependent variable (i.e. all cases in an outbreak are classified as outbreak cases).

Among the models that produced reasonable estimates, the confidence intervals associated with the model using robust variance estimates tended to be equal to or somewhat smaller than those produced by an ordinary logistic model that ignored clustering. Once the clustering effect was accounted for in the variance estimates, the more conservative confidence intervals, that were expected, were achieved (Table 2). The random effects models based on pseudo-likelihood methods worked reasonably well, although, as expected, the model using MQL for linearization tended to have odds ratios closer to the null [30]. Surprisingly, the random effects model using PQL and the weighted logistic regression had unusually similar estimates for the odds ratios and confidence intervals of the significant variables in the final model (Table 2). These values also fell within the ranges predicted using a technique that only included one case per outbreak. Our attempt to weight cases by the inverse of the number of cases associated with an outbreak appears to be a valid approach based on these results. Unlike random effects models, this technique does not provide variance estimates for the different levels in the hierarchy of the data, however, understanding this structure does not have biological meaning in the context of this study.

The results of our study have both epidemiological and methodological implications. The increased risk for young children to be in outbreaks suggests that more attention needs to be paid to issues that place children at greater risk of exposure to point sources of infection that cause outbreaks. Unfortunately, our inability to find interaction effects between age and mode or site of transmission does not provide a specific activity or area to target. Even with 3 years of data, it is obvious that our power to detect interaction effects may be limited when comparing age to a variable with a large number of categories. In terms of methodology, we have found that using weights to account for the size of an outbreak is a useful and robust approach for dealing with clustering when comparing outbreak (i.e. clustered) to sporadic (i.e. independent) cases.

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

ACKNOWLEDGEMENTS

We thank Ian Dohoo and Henrik Stryhn for providing a number of useful suggestions for analysing our data. D.L.P. has been supported by fellowships and awards from the Canadian Institutes of Health Research and the Ontario Veterinary College. We also acknowledge the support of the Wellcome Trust through the International Partnership Research Award in Veterinary Epidemiology.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Ochoa TJ, Cleary TG.** Epidemiology and spectrum of disease of *Escherichia coli* O157. *Current Opinion in Infectious Diseases* 2003; **16**: 259–263.
2. **MacDonald DM, *et al.*** *Escherichia coli* O157:H7 outbreak linked to salami, British Columbia, Canada, 1999. *Epidemiology and Infection* 2004; **132**: 283–289.
3. **Hilborn ED, *et al.*** A multistate outbreak of *Escherichia coli* O157:H7 infections associated with consumption of mesclun lettuce. *Archives of Internal Medicine* 1999; **159**: 1758–1764.
4. **Hrudey SE, *et al.*** A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world. *Water Science and Technology* 2003; **47**: 7–14.
5. **Bruce MG, *et al.*** Lake-associated outbreak of *Escherichia coli* O157:H7 in Clark County, Washington, August 1999. *Archives of Pediatrics and Adolescent Medicine* 2003; **157**: 1016–1021.
6. **Galanis E, *et al.*** Investigation of an *E. coli* O157:H7 outbreak in Brooks, Alberta, June–July 2002: the role of occult cases in the spread of infection within a daycare setting. *Canada Communicable Disease Report* 2003; **29**: 21–28.
7. **Pritchard GC, *et al.*** Verocytotoxin-producing *Escherichia coli* 0157 on a farm open to the public: outbreak investigation and longitudinal bacteriological study. *Veterinary Record* 2000; **147**: 259–264.
8. **Howie H, *et al.*** Investigation of an outbreak of *Escherichia coli* O157 infection caused by environmental exposure at a scout camp. *Epidemiology and Infection* 2003; **13**: 1063–1069.
9. **Havelaar AH, *et al.*** Disease burden in The Netherlands due to infections with Shiga toxin-producing *Escherichia coli* O157. *Epidemiology and Infection* 2004; **132**: 467–484.
10. **Dundas S, Todd WT.** Clinical presentation, complications and treatment of infection with verocytotoxin-producing *Escherichia coli*. Challenges for the clinician. *Symposium Series* (*Society for Applied Microbiology*) 2000; **29**: 24S–30S.
11. **Dundas S, *et al.*** The central Scotland *Escherichia coli* O157:H7 outbreak: risk factors for the hemolytic uremic syndrome and death among hospitalized patients. *Clinical Infectious Diseases* 2001; **33**: 923–931.
12. **Mead PS, *et al.*** Food-related illness and death in the United States. *Emerging Infectious Diseases* 1999; **5**: 607–625.
13. **Michel P, *et al.*** Estimation of the under-reporting rate for the surveillance of *Escherichia coli* O157:H7 cases in Ontario, Canada. *Epidemiology and Infection* 2000; **125**: 35–45.
14. **Rangel JM, *et al.*** Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982–2002. *Emerging Infectious Diseases* 2005; **11**: 603–609.
15. **Mead PS, *et al.*** Risk factors for sporadic infection with *Escherichia coli* O157:H7. *Archives of Internal Medicine* 1997; **157**: 204–208.
16. **Locking ME, *et al.*** Risk factors for sporadic cases of *Escherichia coli* O157 infection: the importance of contact with animal excreta. *Epidemiology and Infection* 2001; **127**: 215–220.
17. **Pearl DL, *et al.*** The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli* O157 in space and time in Alberta, Canada (2000–2002). *Epidemiology and Infection* 2006; **134**: 699–711.
18. **Parry SM, Salmon RL.** Sporadic STEC O157 infection: secondary household transmission in Wales. *Emerging Infectious Diseases* 1998; **4**: 657–661.
19. **Schukken YH, *et al.*** Analysis of correlated discrete observations: background, examples and solutions. *Preventive Veterinary Medicine* 2003; **59**: 223–240.
20. **Goldstein H.** *Multilevel Statistical Models*, 3rd edn. London: Arnold, 2003, pp. 253.
21. **Dohoo IR, Martin W, Stryhn H.** *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC Inc., 2003, pp. 706.
22. **Kulldorff M.** Information Management Services. SaTScan v. 3.0: software for the spatial and space-time scan statistics, 2002.
23. **Skrondal A, Rabe-Hesketh S.** Some applications of generalized linear latent and mixed models in epidemiology: repeated measures, measurement error and multilevel modeling. *Norwegian Journal of Epidemiology* 2003; **13**: 265–278.
24. **StataCorp.** *Stata Base Reference Manual*, volume 2 (G-M), release 8. College Station, Texas: Stata Press, 2003, pp. 531.
25. **Crump JA, *et al.*** Outbreaks of *Escherichia coli* O157 infections at multiple county agricultural fairs: a hazard of mixing cattle, concession stands and children. *Epidemiology and Infection* 2003; **131**: 1055–1062.
26. **Rowe PC, *et al.*** Risk of hemolytic uremic syndrome after sporadic *Escherichia coli* O157:H7 infection: results of a Canadian collaborative study. Investigators of the Canadian Pediatric Kidney Disease Research Center. *Journal of Pediatrics* 1998; **132**: 777–782.
27. **Strausbaugh LJ, Sukumar SR, Joseph CL.** Infectious disease outbreaks in nursing homes: an unappreciated

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press

hazard for frail elderly persons. *Clinical Infectious Diseases* 2003; **36**: 870–876.

28. **Duffell E, *et al.*** Investigation of an outbreak of *E. coli* O157 infections associated with a trip to France of schoolchildren from Somerset, England. *Eurosurveillance* 2003; **8**: 81–86.

29. **Fotheringham AS, Brunsdon C, Charlton M.** *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage Publications, 2000, pp. 270.

30. **Rasbash J, *et al.*** *A User's Guide to MLwiN*, version 2.0. London: Institute of Education, University of London, 2004, pp. 256.

https://doi.org/10.1017/S0950268807008904 Published online by Cambridge University Press