ANALYSIS OF CLUSTERED DATA WITH BINARY OUTCOMES : GIARDIA IN DAIRY CATTLE

H. O. Mohammed [1], Susan Wade [1] and M. Sanaa [2]

SUMMARY : *The purpose of this presentation was to illustrate the application of random-effect models to veterinary data and demonstrate the interpretation of the results obtained from such models. We used data that have a hierarchical structure to address the objective of this study. The data used have a two levels of nesting structure where information on risk factors were collected from individual animals that are grouped in herds and the herds were located in watersheds. In the analysis we modelled simultaneously the correlation at the first level of nesting among herds within a watershed and the second level of nesting between cows within a herd. With the use of an appropriate software, we were able to detect a significant intra-group correlation in the data and evaluate the impact of this correlation on the risk estimates.*

RESUME : *L'objectif de cette communication est d'illustrer l'application des modèles à effets aléatoires aux données vétérinaires et d'interpréter les résultats obtenus pour chaque modèle. Nous avons utilisé, pour cette étude, des données ayant une structure hiérarchique. Les données utilisées ont deux niveaux d'emboîtement. Tout d'abord, l'information concernant les facteurs de risque ont été collectées pour chacun des animaux groupés en élevages, eux-mêmes groupés en fonction de la source d'approvisionnement en eau. Pour l'analyse, nous avons modélisé simultanément la corrélation entre les élevages d'une même source d'approvisionnement (premier niveau d'emboîtement) et la corrélation entre les vaches au sein de l'élevage (second niveau d'emboîtement). Avec l'utilisation d'un logiciel approprié, nous avons démontré l'existence d'une corrélation intra-groupe significative et nous avons pu évaluer son impact sur l'estimation des risques.*

I – INTRODUCTION

One of the most perplexing problems facing epidemiologist is to accurately assess and quantify the relation between putative risk factors and disease. In veterinary medicine data are often collected from groups of animals, e.g., herd, flock, stable, and analyzed using ordinary logistic regression for the purpose of evaluating the association between a disease and the hypothesized risk factors. In using this analytical technique researchers assume independence between animals or study units with respect to the disease status in the study population. This assumption is often violated when the study units originate from populations that are comprised of multiple groups (herd, flock, stable) because animals in the same group tend to be similar with respect to the likelihood of the disease. It is not unreasonable to assume that this clustering will lead to a correlation in the likelihood of disease in the study population. This expected correlation between responses occurs because they are dependent on exogenous factors that are associated with these responses, ie, infection with *Giardia spp.*

This correlation between study units in a cluster has the potential to lead to a statistical phenomenon generally referred to as overdispersion. Overdispersion is a condition where by the standard errors of the regression coefficient is underestimated in the traditional logistic regression models [Atwill et *al.*, 1994]. If overdisperssion is ignored, the estimated standard error and the computed test statistics are grossly biased leading to an erroneous conclusion regarding the relative significance of the hypothesized association. Underestimation to up to 3 000 % has been reported in the literature [Rosner, 1982].

[1] College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA
[2] Ecole nationale vétérinaire, 7 avenue du Général de Gaulle, 94704 Maisons-Alfort Cedex, France

Conditioning on observed set of these factors by controlling for their effect in the analysis and including them as covariates in the logistic regression analysis will sometimes achieve approximate conditional independence. However, more often that this correlation in the response arises from both observed and unobserved risk factors. One approach to handle this correlation among the study units is to assume that the unobserved risk factors were randomly distributed among farms and evaluate the overall significance of this assumption by using a mixed-effect logistic regression model.

Giardia lamblia is a parasite responsible for causing the gastrointestinal in man and animals [Craun et *al.*, 1987, 1990]. The disease is characterized by severe gastroenteritis and diarrhea and symptoms may include diarrhea, abdominal pain, and even fever and vomiting [Snodgrass et *al.*, 1986]. The organisms can be transmitted through water and have been found to be prevalent in unfinished water sources. *Giardia spp.* can survive modern drinking water treatments have been implicated in several drinking water outbreaks throughout the world, including North America and the United Kingdom.

The objectives of this study was to demonstrate how clustered data can be analyzed, estimate of the intra-group association can be obtained, the relationship between the conditional and marginal logistic regression, and how the results of the analysis can be interpreted. These objectives will be addressed through an example of a cross-sectional study of Giardia in dairy cattle. This study was carried out to determine prevalence of *Giardia* in dairy cattle in a specific population and to identify factors associated with the prevalence.

II - MATERIALS AND METHODS

II.1. STUDY POPULATION

The target population consisted of dairy herds in four counties in southeastern New York state. This target population was selected because they are located in New York city watershed. A random sample of dairy herds was selected from the target population. A detail description of the target population and the sampling procedure were provided earlier [Wade et *al.*, 1998]. The sampling list-frame consisted of dairy herds in that geographic area and the sampling units was the dairy. A proportional sampling scheme was adopted where sample were collected from : all calves below 6 month, 6 heifers that were > 6 month to first freshening, and 9 cows. The fecal samples were analyzed for the presence of *Giardia spp.* using a modification of the floatation method [George, Wade et *al.*, 1997].

II.2. DATA COLLECTION

Several management and health factors were hypothesized to associate with the risk of infection with *Giardia sp.* A questionnaire was designed and validated to collect data on these hypothesized factors. The factors included intrinsis, demographic, and management factors. A detailed description of the management factors investigated in this study is provided elsewhere [Wade et *al.*, 1998].

II.3. DATA ANALYSIS

Furthermore, management and intrinsic factors significantly associated with the risk of infection with *Giardia spp.* were evaluated using the logistic regression analysis with two-step backward elimination. These factors included the four indices, the age of the animal and the season of sampling. The probability that an animal was infected with *Giardia spp.* given a set of factors was evaluated from the logistic regression analysis as follows :

$$P(Gij/\alpha,\beta_i) = \frac{1}{.\ (1+\exp^{-(\alpha+\Sigma\beta k\, Z_i)})}$$

where P(G) is the probability that the animal become infected with *Giardia spp.*, α is the log odds of infection with *Giardia spp* under standard circumstances, and β_i is the change in the log odds of infection due to the presence of a particular factor Z_i.

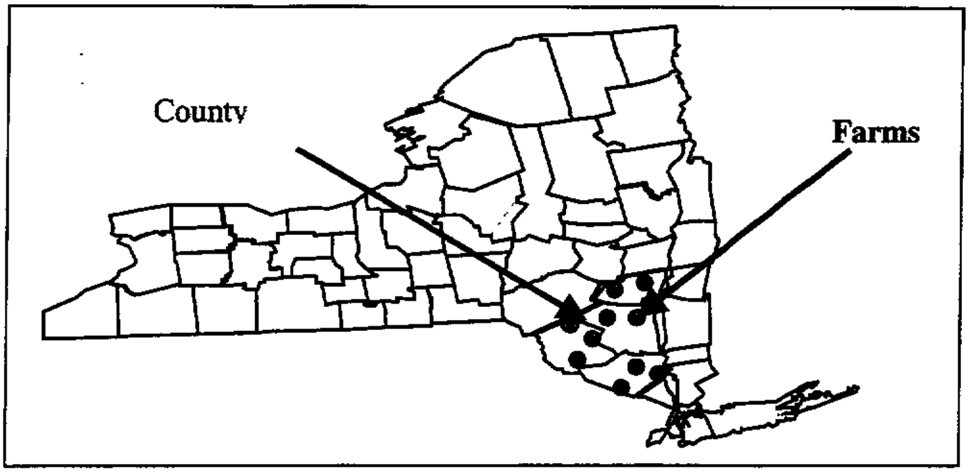
Step 3. As we stated earlier because the sampling units, the animal, in this study are clustered in herds or farms, and the farms are clustered within a watershed we hypothesized that this clustering will lead to a correlation in the likelihood of infection with *Giardia spp.* in the study population (figure 1). This expected correlation between responses occur because they are dependent on exogenous factors that are associated with these responses, ie, infection with *Giardia spp.* Conditioning on observed set of these factors by controlling for their effect in the analysis and including them as covariates in the logistic regression analysis will sometimes achieve approximate conditional independence. However, more often that this correlation in the response arises from both observed and unobserved risk factors. We assumed that the unobserved risk factors were randomly distributed among farms and we evaluated the overall significance of this assumption by using a mixed-effect logistic regression model. A random effect term (μ) to capture the herds' variability was added to the logistic regression model and evaluated for its significance using the likelihood ratio statistics. This mixed effect logistic regression model for the herd effect was specified as follows :

$$P(Gij/\alpha,\beta_i,\sigma) = \frac{1}{(1+\exp^{-(\alpha+\Sigma\beta k\, Z_i+\mu_i\sigma)})}$$

where $P(G_{ij}/\alpha, \beta_k, \sigma)$ is the probability that an animal within a herd level α_i would have been infected with *Giardia spp.* given standard set of fixed factors Z_{ijk} with an effect of β_k, and potential unmeasured herds' (μ_i) factors with their estimated standard error, σ. The modified likelihood ratio test was used to test the null hypothesis that σ is equal to zero ($H_0 : \sigma = 0$). The mixed effect logistic regression analysis was performed using the MIXOR statistical software.

FIGURE 1

The conceptual structure of the data showing hierarchical relationship



Since the herds were located in different watershed, we assumed that these herds were nested within watershed and this relation affects the likelihood of infection within Giardia. We first tested the first level, watershed, random effect by including a term, $v_{watershed}\sigma_1$ to evaluate whether the infection status of the animals was correlated by watershed. Then we extended the model by adding a second term, $\mu_{herd}\sigma_2$, to test wheter there was additional intra-herd correlation with respect to the infection status if we control for the potential correlation at the watershed level. The nested model was specified as follows :

$$P(Gij/\alpha,\beta_i,\sigma_1,\sigma_2) = \frac{1}{(1+\exp^{-(\alpha+\Sigma\beta k Z_i+v_i\sigma_1+\mu_i\sigma_2)})}$$

where $P(G_{ijk}/\alpha,\beta_l,\sigma 1,\sigma 2)$ was extended to include both random effects, $v_{watershed}$ and μ_{herd}. The signifcance of the random effects were evaluated using the modified likelihood ratio test in MIXOR. The characteristics of this model that the log odds ratio for a specific risk factor is conditional on the random effect, the odds ratios in each model are group (herd or watershed) specific, and that the estimates of these odds ratios vary from group to group.

In cases where the modified likelihood ratio test was significant, we concluded that there was evidence for correlation among animals in relation to the infection status with Giardia. We estimated the intra-group correlation (ρ) on the logit scale from the random effect variance estimates as follows :

$$\rho=\frac{\sigma^2_{RE}}{\sigma^2_{RE}+\sigma^2_{ST}}$$

Where σ^2_{RE} is the variance of the random effect obtained from the mixed-effect model, σ^2_{ST} is the fixed variance of the standard logistic regression model and equal $\pi^2/3 = 3.3$ [Searle et *al.*, 1992]. This ρ is interpreted as the total variation for the infection status on the logit scale that is attributed to the grouping variable (herd or watershed). From the above equation we notice that $\rho \approx 0$ if $\sigma^2_{RE} \approx 0$ and $\rho \approx 1.0$ when σ^2_{RE} is significantly larger than 3.3.

If the nested effect was significant, the inter class correlation was computed as follows :

$$\rho=\frac{\sigma^2 RE1+\sigma^2 RE2}{\sigma^2 RE1+\sigma^2 RE2+\sigma^2 ST}$$

Where σ^2_{RE1}, σ^2_{RE2} are the estimated variance for the two levels of random effects, watershed and herd. Estimates for these two parameters were obtained from the mixed effect model.

Estimates of the conditional coefficients in the mixed effect model can be obtained from the estimated coefficients in the standard logistic regression model. The relation between the two estimates is determined as follows :

$$\beta C = \beta M \times \sqrt{1+\frac{\sigma^2 RE_1+\sigma^2 RE_2}{2.89}}$$

Where the β_C is the effect as estimated in the conditional logistic regression model, β_{MS} are the marginal coefficients estimated in the ordinary logistic regression model. The relation between the marginal and conditional coefficients is described in detail in Zeger et *al.* [1988].

III - RESULTS

A total of 2 941 animals on 109 farms were investigated in this study. Thirteen percent of the animals were found to be infected with *Giardia spp.* The organism was detected in fecal samples collected from animals of all age groups with animals below 6 months of age represents the majority.

Several management factors were found to significantly associate with the risk of infection with Giardia in the ordinary logistic regression analysis (table I). Factors associated with increased risk of infection with *Giardia spp.* Included use of mechanical ventilation, less frequent addition of bedding, use of saw dust for bedding, and use of ionophores as prophylactics. Factors associated with decreased risk of infection included the use of milk replacers and use of antibiotics as prophylactics.

TABLE I

Factors found to significantly associate with the risk of Giardia spp. Infection in the ordinary logistic regression analysis

RISK FACTOR	REGRESSION COEFFICIENT	STANDARD ERROR	90 % CONFIDENCE INTERVAL
Type of ventilation			
Natural	0.0		
Mechanical	0.608	0.275	(1.2 – 2.9)
Use of milk replacer			
No	0.0		
Yes	-0.446	0.205	(0.4 – 0.9)
Addition of bedding			
Daily	0.0		
Less frequently	0.801	0.317	(1.3 – 3.8)
Type of bedding			
Straw	0.0		
Saw dust	0.704	0.202	(1.5 – 2.8)
other	0.303	0.303	(0.9 – 1.9)
Use of antibiotics as prophylactics			
No	0.0		
Yes	-0.348	0.158	0.5 – 0.9
Use of ionophores as prophylactics			
No	0.0		
Yes	0.352	0.156	1.1 – 1.8
Constant	-2.150	0.346	

Table II shows the results of the nested mixed effect analysis after adjusting for the watershed and herd effects. There was a significant correlation between animals in relation to their risk of infection with *Giardia spp.* that were associated with the watershed and the herd in which the animal originated This conclusion was made because the random effect parameters for the watershed and herd were significant. The interpretation of this correlation can be best described in figure 2. Instead of having one logistic regression curve for the probability of infection with *Giardia spp.* For the study population, we will have several curves each representing a population depending on the amount of intra-group correlation between individual animals in the herd.

As a result of this intra-groups correlation the standard error for the estimated effect (regression coefficient) for the use of milk replacers and ionophores were higher under the mixed effect models in comparison to the logistic regression model. Figure 3 shows the variability in the intra-group correlation, ρ, among animals in this study population. The value of ρ varied from 0.42 to 0.27.

TABLE II

Factors found to significantly associate with the risk of Giardia spp. Infection in the ordianry logistic regression analysis

RISK FACTOR	ORDINARY		MIXED EFFECT	
	REGRESSION COEFFICIENT	STANDARD ERROR	REGRESSION COEFFICIENT	STANDARD ERROR
Type of ventilation				
Natural	0.0		0.0	
Mechanical	0.608	0.275	0.608	0.285
Use of milk replacer				
No	0.0		0.0	
Yes	-0.446	0.205	-0.446	0.344
Addition of bedding				
Daily	0.0		0.0	
Less frequently	0.801	0.317	0.801	0.319
Type of bedding				
Straw	0.0		0.0	
Saw dust	0.704	0.202	0.704	0.212
other	0.303	0.215	0.303	0.429
Use of antibiotics as prophylactics				
No	0.0		0.0	
Yes	-0.348	0.158	-3.49	0.155
Use of ionophores as prophylactics				
No	0.0		0.0	
Yes	0.352	0.156	0.352	0.421
Constant	-2.150	0.346	-2.150	0.547
Scaler watershed			0.063	0.030
Scaler farm			0.820	0.234

FIGURE 2

Conceptual presentation of the probability of the disease in the study population because of the significance of the random effect parameters

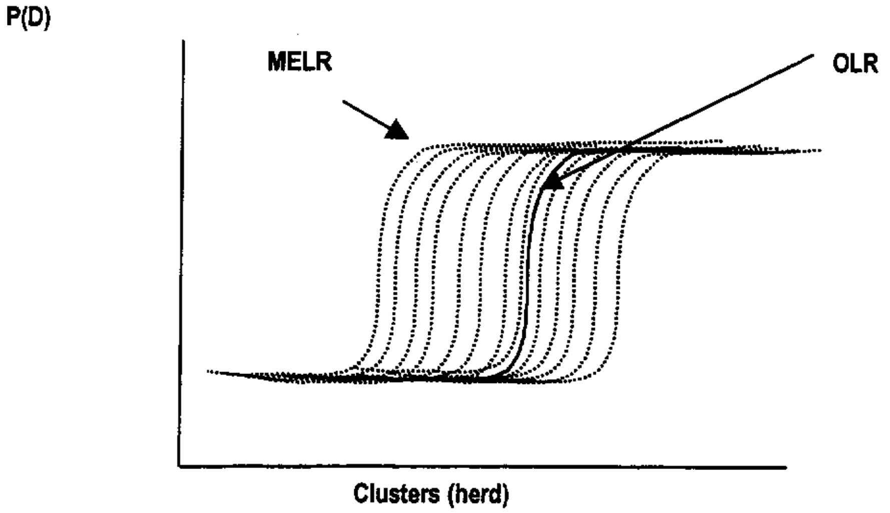
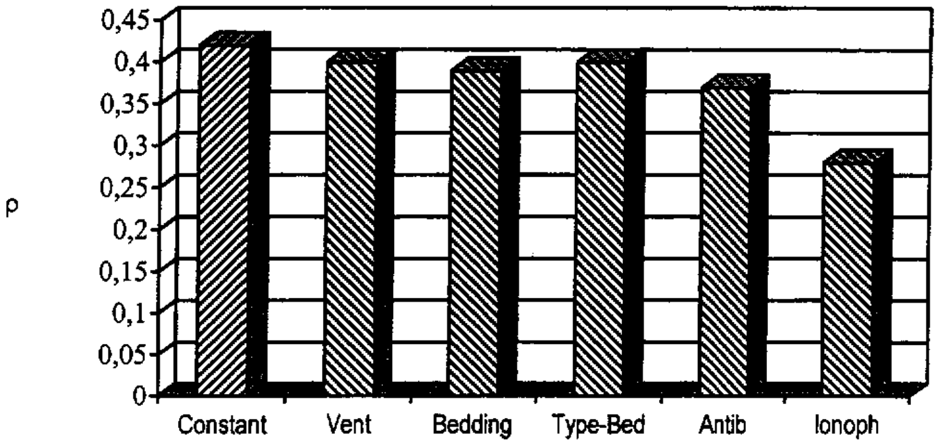


FIGURE 3

The intra-group correlation (ρ) among the study population



IV - DISCUSSION

Ordinary logistic regression has become the method of choice in veterinary epidemiology when an investigator wants to regress one or more predictors or risk factors on a binary out come such as disease status, positive or negative. The theory behind the use of this model requires that the disease status of each study unit to be independent from other animals in the study population. However, the majority of the epidemiologic studies in the veterinary literature has some level of dependency and this dependency should be taken into consideration when performing analysis. Methods that ignoring the hierarchical or nested structure of the data are likely to under-estimate the variance of the estimated effect of a factor(s). Moreover, in the ordinary logistic regression where the underlying relationship between the putative factors and the disease of interest is not linear, failure to account for the clustering of the data has the potential to result in reproducible parameter estimates.

One possible approach to examine for this dependency is to use the fixed effect models to account for this variation by cluster or a group. However, this approach is not practical if the is a large number of clusters or groups in the study. Furthermore, the confidence interval and the test statistics are based on asymptotic normality of the estimate and asymptotic χ^2 distribution for the likelihood ratio test, these two breakdown when the number of parameters to be estimated is large.

A more powerful solution is to assume a finite set of levels of factors that are hypothesized to associate with the disease under investigation and treat them as covariates. In addition, make the assumption that the random effect is attributed to an infinite set of level of factors from which a random sample is drawn in the study. The mixed-effect models gets around the problem by estimating the variance of distribution of the group effect rather than separate parameter for each cluster or group.

There are other advantages for using the random effect models for grouped or clustered data. The statistical techniques used in data analysis assume that there is no measurement error in the estimation of the specific value of a predictor or a disease status. However, we all know that most of the measurement tools are not perfect, have a 100 %sensitivity and specificity. By using the random effect analysis approach, one can assume that the measurement error is non-differential in nature and randomly distributed among the study units. This assumption can be tested through the use of random effect models.

The use of mixed effect models in addition to providing accurate estimate of the relation between a putative risk factor and a disease, also yield important information in relation to the similarities between individuals within a cluster. By estimating the intra-group correlation one could get an important insight about the influence of the group level effect on an individual's risk of disease. Such knowledge is important to device effect intervention strategies.

Another question of interest that should be entertained before making a decision regarding the analytical approach to employ is the cost of ignoring the multilevel structure of the data. Consideration should be given to the potential bias that might occur in the estimates of the relation between the putative risk factors and the disease under investigation. It requires knowledge of advance statistical techniques to perform such analyses and also the acquirement of specialized software. These costs have to be weighed against the potential biases of ignoring the multilevel nature of the data at hand. An argument could be made to whether one would ever be able to measure bias, however, the use of simulated data with Monte Carlo estimates might provide a reasonable altemative to evaluate bias.

We have detected a significant intra-group correlation among in our study population. Several factors contribute to this correlation which we did not account for in this study. Although it is difficult to pinpoint the source of this correlation, one could speculate on some of the factors that attribute to its existence. One obvious factor is the nature of the disease, the more contagious the disease under study the more similar are the study units [Rosner, 1989]. Also, if the management practices to control the disease are similar it is also likely to have a similar responses to the disease in the population. This intra-group correlation provides more insight into the influence of the group on the risk of the disease.

V – CONCLUSION

We have demonstrated in this study that the application of random effect models to hierarchical data with a binary response. Our findings have emphasized the importance of evaluating data for the intra-group correlation and quantifying this correlation. Ignoring this intra-group correlation in such a data is likely to lead to information bias in relation to estimating the significance of association of risk factors and likelihood of a disease. In estimating any such relationship between a factor and a response it is important to keep in mind that the posterior probability is not a point estimate but it does vary by group or cluster and by the amount of intra-group correlation in each group.

VI - REFERENCE

Atwill E.R., Mohammed H.O., Scarlett JM, McCulloch CE. ~ Extending the interpretation and utlity of the mixed effect logistic regression models. *Prev. Vet. Med.* 1995; **24**, 187-202.

Craun G.F., Jakubowski W. ~ Status of waterborne giardiasis outbreaks and monitoring methods. *In*: Proceedings of the International Symposium on Water Related Health Issues, CL Tate Jr, ed. Amer Water Resources Assoc, Bethesda, MD. 1987;167-174.

Craun G.F. ~ Waterborne giardiasis. chap 15. *In*: Giardiasis, EA Meyer, ed. Elsevier, Amsterdam, 1990, 267-293.

Dixon W.J. ~ BMDP Statistical Software Manual. Page 1105, Univ Calif Press, Berkeley, CA., 1990, 1105-1144.

Georgi J.R., Georgi M.E. ~ *Parasitology for Veterinarians*, 5th ed. Philadelphia, PA; WB Saunders Co, 1990.

Rosner B. ~ Multivariate methods for clustered binary data with more than one level of nesting. *J. Am Stat. Assoc.*, 1989, **84**, 373-380.

Snodgrass,D. R., Terzoro H.R., Sherwood D., Campbell I., Menzies J. D., Synge B. A. ~ Aetiology of diarrhea in young calves. *Vet Rec.*, 1986, **119**, 31-34.

Stiratelli R., Laird N.M and Ware J.H. ~ Random-effect models for serial observations with binary response. *Biometrics*, 1984, **4**, 961-971.

Wade S.E., Mohammed H.O.,. Schaaf S.L. ~ Prevalence of *Giardia* sp., *Cryptosporidium parvum* and *Cryptosporidium muris* in 109 Dairy Herds in Five Counties of Southeastern New York. Am. J. Vet. Med. Assoc. 1998 (Submitted).

Zeger S. L., Liang K -Y. and Albert P. S. ~ Models for longitudinal data : A generalized estimate equation approach. *Biometrics*, 1988, **44**, 1049-1