2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Herd-level Risk Factors Associated with Bovine Brucellosis Seropositivity and Abortion in Bangladesh

Promotor :
Prof. dr. Marc AERTS

Promotor :
Prof.dr. DIRK BERKVENS

Md. Atiqul Islam

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Herd-level Risk Factors Associated with Bovine Brucellosis Seropositivity and Abortion in Bangladesh

Promotor :
Prof. dr. Marc AERTS

Promotor :
Prof.dr. DIRK BERKVENS

Md. Atiqul Islam

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Certification

I declare that this thesis was written by me under the guidance and counsel of my supervisors.

…….. …… Date……
Md. Atiqul Islam Student

We certify that this is the true thesis report written by **Md. Atiqul Islam** under our supervision and we thus permit its presentation for assessment.

…….. …… Date……
Prof. dr. Marc Aerts Internal Supervisor

…….. …… Date……
Prof. dr. Dirk Berkvens External Supervisor

Dedicated to

My

Beloved Parents

Acknowledgements

It is a great opportunity to convey my deepest indebtedness to my supervisors Prof. dr. Marc Aerts, Universiteit Hasselt, Diepenbeek, Belgium and Prof. dr. Dirk Berkvens, Institute of Tropical Medicine, Antwerpen, Belgium for their constant encouragement and invaluable personal guidance for carrying out my thesis. I express my deep feelings and gratitude to them not only for supervising my work but also for many more helpful suggestions they extended me without which it would have not been possible for me to complete my research.

I express my profound acknowledgement to the Vlaamse Interuniversitaire Raad (VLIR) for granting me the scholarship in order to fulfill my higher study in Universiteit Hasselt.

I would like to thank AKM Anisur Rahman and dr. Abatih Emmanuel for helping me the understanding of the technical terms of animal studies, sampling techniques and many things in this research. I am also grateful to all of my respected teachers and staffs of the Censtat, UHasselt and the department of Statistics, SUST, Sylhet, Bangladesh. I desire to express my especial thanks to Patwary, Nolen, Rasheda, Veronica, Fitsum, Samson, Emmanuel, Chellafe and many more for their co-operation during the entire two years of my study in Belgium.

On a more personal level, I would like to express my heartiest honour to my parents and uncle Md. Ataur Rahman whose affections and encouragement enabled me to finish this research work. I am also thankful to my family members, especially my brothers (Robi & Rifat), sisters (Bonani, Salma & Rani), nephew and niece for their patience and understanding during the period of my study in UHasselt, Belgium.

I am solely responsible for errors and omissions in this dissertation, if any.

Md. Atiqul Islam

September 12, 2011

Abstract

Correlated data are common in veterinary epidemiology, where clustered and hierarchical data are often studied. A cross-sectional study was conducted to determine the seroprevalence and to identify the herd level risk factors associated with bovine brucellosis seropositivity and abortion in the Mymensingh and Sherpur districts of Bangladesh. A total of 388 herds were selected from the 29 unions. Generally, herds are clustered within unions (areas) and therefore the probability of herds within unions being more similar than between unions cannot be ignored. A herd was considered seropositive if at least one animal was tested positive within the herd by any of the three tests (RBT, SAT or iELISA). The overall herd level prevalence of bovine brucellosis and abortion were 35.10% and 20.62% respectively. Due to the clustered nature of the data, different techniques were employed for analysis. A random effects approach was used to account for the correlation in the data allowing us to study both the population-averaged and subject-specific (union-specific) models. Using generalized estimating equations (GEEs), herdsize and interaction between breeding and herdsize were the important risk factors associated with both brucellosis seropositivity and abortion. Since the cluster size (unionsize) was informative in the case of abortion, cluster-weighted generalized estimating equation (CWGEE) was also employed. For the union-specific risk factors, generalized linear mixed model (GLMM) was also used. In addition, a joint random intercepts model was fitted to identify the association between brucellosis seropositivity and abortion. The association however was not statistically significant. Nevertheless, from all the analyses, breeding, herdsize and their interaction were the main herd level risk factors associated with bovine brucellosis seropositivity and abortion. Thus, the brucellosis control programs will be beneficial if these risk factors are taken into account.

Keywords: Abortion, Bovine Brucellosis, Breeding, Cluster-weighted GEE, Joint Random Intercepts Model

Contents

List of Tables

List of Figures

1. Introduction

Worldwide, Brucellosis is a zoonotic disease caused by gram-negative bacteria *Brucella* that are pathogenic for a wide variety of animals and human beings (Matyas and Fujikura, 1984). It is mainly a disease of domestic animals (e.g. cattle, buffalo, sheep, goat, dog, and pig) caused by various species of *Brucella* viz. *B. abortus*, *B. melitensis, B. ovis, B. canis, B. suis*. The greatest economic impact results from bovine brucellosis usually caused by biovars of *B. abortus*. In some countries however, mainly in southern Europe and western Asia, where cattle are kept in close association with sheep or goats, infection can also be caused by *B. melitensis* (Young, 1995; OIE, 2000).

Brucellosis is still wide spread and its prevalence is increasing even with the advances made in the diagnosis and control especially in many developing countries where rural income depends largely on livestock breeding and dairy products (Vandeplassche, 1982; Roth *et al.,* 2003). In cattle, infection causes herd production losses as a result of abortions, reduction in milk production, increasing calving intervals, the birth of weak calves and increased death rates due to metritis, following retention of the placenta (Vandeplassche, 1982; Blood and Radostits, 1989; Stringer *et al.,* 2008). Cattle abortions due to *Brucella* usually take place at between six and eight months of gestation. *Brucella* and other infections are suspected for specific causal factors of abortion in cattle though they have not been fully specified (Vandeplassche, 1982).

Generally, the diseases of animal populations are studied scientifically in veterinary epidemiology. Its aim is to quantify the disease and also identify the risk factors that may have an effect on the occurrence of the disease. This information is then used to prevent or reduce the extent of the problem or disease. Quantification of the disease is often based on the prevalence, describing the probability that an animal from the population has the disease (Faes *et al.,* 2006). In this study, we are confronted only with the herd level risk factors associated with brucellosis seropositivity and abortion in Bangladesh.

Clustered binary data occur commonly in both biomedical and health sciences or even veterinary epidemiology. The clustering may arise in this study due to sampling of the primary sampling unit (herd), for instance, when observations are made on each member within a cluster (union) or group. Whatever the nature of the clustering, observations (herds) within the same cluster (*union*) is usually correlated (Fitzmaurice, 1995).

Like many other infectious disease data, this study was also confronted with the problem of clustering in the dataset. At the animal level, once an infected animal is introduced into a herd, others animals within the herd have an increased instantaneous risk of becoming infected. At the herd level, if one of the animals within the herd is infected then the whole herd is treated as infected. So, if brucellosis positive animals are introduced into a *Union* (Area), herds within the same union might have a higher risk of becoming infected. Thus, individual responses are more homogeneously distributed within herds/unions than across herds/unions. In modeling such studies, it is a good practice to work with models that take into account the clustering effect (Aerts *et al.,* 2002). There are several ways to deal with such clustering, some of which estimate marginal, population-averaged measures of effect and some of which estimate the subject-specific measures of effect, e.g. random-effects models. This study also presents the methods to derive population-averaged and union-specific risk factors of brucellosis seropositivity and abortion based on generalized estimating equations (GEEs) and generalized linear mixed model (GLMM) (Molenberghs and Verbeke, 2005). It can be added here that this is the standard situation for the analysis of univariate clustered data. When multivariate longitudinal or clustered data arise, instead of a single outcome, a set of different outcomes on the same unit is measured repeatedly over time or for subjects nested within naturally occurring groups. These outcomes may be of similar or disparate types, and a variety of scientific questions may be of interest, depending on the application (Fitzmaurice *et al.,* 2008). If more than one outcome is present, a mixed (or generalized linear mixed) model can be used for each. These separate models can be tied together into a multivariate mixed (or generalized linear mixed) model by specifying a joint distribution for their random effects. This strategy has already been used for joining multivariate longitudinal profiles or other types of multivariate clustered data (Fieuws *et. al.,* 2007).

1.1 Review of Literature

A review of studies related to the present study reveals risk factors which may affect the occurrence of brucellosis seropositivity and abortion. This review concentrates around some of the pioneer works extensively used in seroprevalence of brucellosis as well as abortion of animal and herd level in the context of Bangladesh as well as other countries; hence an overview of the previous relevant works is presented in this section.

In Bangladesh, brucellosis was reported by Mia and Islam (1967) that 37% of adult cows were infertile and it may play an important role in causing infertility in cows. The estimated annual economic loss due to bovine brucellosis in indigenous cows in Bangladesh was 720,000 EUR (in total) and 12,000 EUR per 1,000 cross-bred cows (Islam *et al*., 1983). Pharo *et al*. (1981) estimated herd level prevalence of bovine brucellosis as 62.5% in the Pabna milk-shed area of Bangladesh by using milk ring test (MRT). About 30.7% of MRT positive cows were found to be RBPT (Rose Bengal plate test) positive. Rahman and Rahman (1982) carried out a study on the prevalence of brucellosis in cows in organized farms and domestic holdings in Bangladesh. It was observed that 8.47%, 1.63% and 0.41% sera samples of the MRT positive cows from pabna, Faridpur and Bogra districts respectively were also RBPT posiitve. Rahman *et al.* (2006) reported the animal-level seroprevalence of brucellosis in cattle as 2.4%-18.4%. A cross-sectional study was conducted by Nahar and Ahmed (2009) and reported an overall seroprevalence of brucellosis in cattle as 4.5% and the prevalence and risk factors of brucellosis were greatly influenced by age, gender, breed, area, pregnancy status and grazing pattern in cattle. Amin *et al.* (2004) carried out a serological survey of bovine brucellosis in cows of *Mymensingh* district of Bangladesh. The highest prevalence was recorded in cows above four years of age, with a history of previous abortion, in repeat breeders and in retention of placenta. A recent cross-sectional study was conducted by Ahasan *et al.* (2010) to determine the seroprevalence and potential risk factors of brucellosis in cattle in Dinajpur and *Mymensingh* districts of Bangladesh. Cattle were examined by Rose Bengal Plate Test (RBPT) and later confirmed positive cases by Serum Agglutination Test (SAT) and both indirect and competitive Enzyme Linked Immunosorbent Assay (iELISA and cELISA). The overall animal level prevalence was 3.30%. Brucellosis seroprevalence was higher in female cows above 48 months

than those of male under 48 months; cattle breed naturally than artificial insemination, with previous abortion record than non-aborted respectively.

A random survey was carried out by Aulakh *et al.* (2008) to study the epidemiology of brucellosis in Punjab of India. The overall apparent prevalence of brucellosis was found to be 18.26%. The prevalence in the central zone of the state was significantly higher, viz. 23.2% (chi square = 11.34, $p < 0.01$) compared to 14.2% in the sub-mountainous zone and 5.8% in the arid irrigated zone. It was found that there was significant association between disease and abortion (chi square = 22.322, $p < 0.01$) and maximum abortion cases due to brucellosis were found in above six months of gestation (95.7%). The disease was significantly associated with the retention of placenta but no significant relationship of the disease with repeat breeding.

Stringer *et al.* (2008) carried out a cohort study to quantify the risk of seropositivity in bovine animals moved from herds infected with brucellosis. It was found from the multivariate logistic regression model that factors influencing the risk of seropositivity in the exposed cohort of animals were maternal status (whether the dam had been a brucellosis reactor) and age at leaving the infected herd. Another cross-sectional study was conducted to identify risk factors for herd infection by *Brucella* spp. in dairy cattle in the suburbs of Asmara, Eritrea. A seropositive herd was defined as one in which at least one animal tested positive in the complement-fixation test (CFT). Mixed-breed herds, compared to single (exotic)-breed herds, were found to be independently associated with increased herd seroprevalence (OR=5.2; 95% C.I: 1.4±18.7) in the multiple logistic model with the herd infection status as the dependent variable. The importance of this variable was supported by the multiple beta-binomial regression model (OR= 3.3; C.I: 1.4±7.6) with animal-level prevalence within herd as the outcome variable. Both models also revealed the presence of a negative association between seropositivity and cattle stocking density (Omer *et al.,* 2000).

Muma *et al.* (2007) reported at the individual animal level that the presence of high levels of anti-*Brucella* antibodies and age of the animal had a significant effect on cattle abortions. No relationship between abortions and anti-*Brucella* antibodies was found at the herd level, but the herd size was noted to be associated with the abortion status of the herd. A recent cross-sectional study was conducted by Matope *et al.* (2010) to investigate factors for *Brucella* seropositivity in

smallholder dairy cattle herds from different agro-ecological regions of Zimbabwe. The herd-level factors for *Brucella* seropositivity were tested using multivariable logistic model with herd infection status as dependent variable while the levels of exposure in individual animals within-herds were analyzed by negative binomial regression using the number of positive animals as the outcome. Using the logistic regression model they identified area, with both Rusitu (OR = 0.26; 95% C.I: 0.07, 1.03) and Wedza (OR = 0.07; 95% C.I: 0.01, 0.49) having lower *Brucella* seropositivity compared to Gokwe. Keeping mixed cattle breeds (OR = 8.33; 95% C.I: 2.70, 25.72) compared to single breed herds, was associated with increased herd seropositivity. The odds of *Brucella* seropositivity were progressively higher with increasing stocking density and herd size. Using the negative binomial regression model they identified area, keeping mixed breed herds, stocking density and herd size as significantly associated with increased counts of seropositive cattle in a herd. A more recent study by Matope *et al.* (2011) used generalized estimating equations and logistic regression to identify the risk factors for *Brucella* spp infection. For herd level, *Brucella* seropositivity, geographical area, purchase of cattle and large herd size were independently associated with increased odds of abortion. Exposure to *Brucella* had a significant impact on abortion.

Despite the above different views there is scant information about the animal and herd level prevalence and risk factors of brucellosis in Bangladesh context using an appropriate study design. Therefore, the present study was carried out to determine the seroprevalence and risk factors of brucellosis and abortion in the herds within the union with the aim to initiate the bovine brucellosis control program.

1.2 Objective

The main objective of the study is to identify the herd level risk factors associated with the prevalence of bovine brucellosis seropositivity and abortion. To cover the main objective, the specific objectives of this study are:

- To investigate the prevalence of bovine brucellosis seropositivity and abortion in *Mymensingh* and *Sherpur* districts.
- To study the union specific prevalence of herd level risk factors associated with brucellosis seropositivity and abortion.

- To study the joint binary responses (*herdposit* and *abortion*) and see the association between the responses and to test whether there are significant effect of risk factors on either abortion or brucellosis seropositivity (*herdposit*).

1.3 Organization of the Study

Following the introduction in Section 1, Section 2 deals with study area, study design, study variables and methodology. The results of the study are presented in Section 3 and finally Section 4 comprises the discussion of findings and end with some concluding remarks on the basis of findings. Furthermore, content has been given at the beginning and the list of the references is given at the end of the study.

2. Data and Methodology

The significance of any research depends on using a reliable source of data. This section provides a brief description of the study area, study design, analysis plan and other related issues of the study.

2.1 Profile of the Study Area

Bangladesh is an irregularly shaped and low-lying country with a total area of 147,570 square kilometers and about 142.319 million people. For administrative purposes, the country is divided into seven divisions, 64 districts, and 500 upazillas (sub-districts) (BBS, 2011). The present study was carried out in *Mymensingh* and *Sherpur* districts, most dense livestock regions in the Dhaka division of Bangladesh located between latitudes 23° 46′ and 25° 12′N and longitudes 90° 04′ and 90° 34′E. The areas were chosen because of the location of Bangladesh Agricultural University (BAU) which manages the brucellosis diagnostic laboratory and also they have the highest livestock population density (>600/sq.km). Total areas are 4363.48 sq km for *Mymensingh* and 1363.76 sq km for *Sherpur*. The *Mymensingh* district consists of 8 municipalities, 12 upazillas and 146 unions whereas *Sherpur* district has one municipality, 5 upazillas and 51 unions. The soil formation of these districts is flood plain, grey piedmont, hill brown and terrace. There are small valleys between the high forests with annual average temperature maximum 33.3°C, minimum 12°C and annual rainfall 2174 mm (Banglapedia, 2011).

2.2 Study Design

A cross-sectional study was carried out to investigate the herd level seroprevalence of bovine brucellosis in the *Mymensingh* and *Sherpur* districts of Bangladesh. The study was conducted between September 2007 and August 2008. Since there is no livestock databank in Bangladesh, the first step of the sampling process was the digitization of the map of *Mymensingh* and *Sherpur* districts using Arc View Version 1.0 (Environmental Systems Research Institute, Inc. Redlands, California). Out of 146 unions (sub *Upazilla*) of *Mymensingh* district (consist of 12 *Upazillas*), 28 were randomly selected. Similarly, one union from *Sherpur* district was also selected. Usually one geographical coordinate was selected randomly from each selected union and located by a hand held GPS reader. Livestock farmers within 0.5 km of the selected point were informed about the survey (Cringoli *et al.,* 2002). To encourage livestock farmers to participate, free anthelmintics and vitamin-mineral premix were supplied to their animals when sampling took place. Finally, 388 herds were selected within the 29 unions depending on the selected area.

2.2.1 Serological procedures

Serum samples were collected from individual animals within the selected herd and tested using the Rose Bengal Plate Test (RBT), a Serum Agglutination Test (SAT) and an indirect Enzyme-linked Immunosorbent Assay (iELISA) respectively. An animal was considered to be seropositive if it returned positive results using any of the three tests. This interpretation of being seropositive of brucellosis is called parallel interpretation. While a serial interpretation of brucellosis seropositive in which three tests together regarded as *Brucella* seropositive. A herd was treated as brucellosis seropositive if at least one animal within the herd tested positive on either test (RBT, SAT or iELISA). The iELISA has the best test characteristics with a sensitivity and specificity of 94.7% and 93.2% respectively. Both RBT and SAT were highly specific (99.5%). The sensitivities of RBT and SAT were very low (34.8% and 29.8% respectively) (Nielsen, 2002).

2.2.2 Epidemiological data collection

A pre-tested structured questionnaire with mostly closed-ended (categorical) questions was used to collect information on animal and herd level risk factors which might be associated with brucellosis and abortion. Pre-testing of the questionnaire was carried out in one of the study areas by interviewing a few farm owners and any lack of clarity of questions was noted and later revised. Since the study only deals with herd level characteristics, so the possible potential herd level risk factors are presented in Table 1. A herd was regarded as positive for abortion if at least one cow within the herd was reported to have a previous record of abortion.

2.3 Variables Description

In this study, some covariates listed in Table 1 were used to explain the response variables.

Table 1: Description of the variables (with coding) in the study of bovine brucellosis

Variable	***Description***
Herdposit (response)	The herd is infected or not (according to iELISA, RBT, SAT tests)
herdid	Identification number of herd
Union	Union (sub-sub-district) in *Mymensingh* and *Sherpur* districts
Abortion (response)	Herd had abortion or not (Yes/No)
Breeding	Artificial Insemination (Yes) or Natural Service (No)
Purchase	Herd purchased animals or not (Yes/No)
Herdsize	Size of herd (Number of cattle)
Unionsize	Size of the Union (Number of herds)
Farmtype	Subsistence (1) or Commercial (2)

2.4 Methodology

2.4.1 Flow of Analyses

This section provides the analysis plan and procedure to address the specific objectives. Firstly, data exploration techniques are presented to review the data structure. Secondly, to identify the primary important risk factors, logistic regression analysis is carried out ignoring the clustering effect in the data. On the other hand, there are two basic methods for handling correlated binary data: one is using a population-averaged method, more specifically the generalized estimating equation (GEE) model; another is the generalized linear mixed model (GLMM), a random effects approach. Finally, a joint random-intercepts model is considered to study the relationship between the two binary responses.

2.4.2 Exploratory Data Analysis (EDA)

This fundamental step has been carried out in order to gain better insight into the data set. Simple descriptive statistics (cross-tabulation) and some graphical representation were mainly used to study the association between the response variables and the set of explanatory variables.

2.4.3 Statistical Analysis

2.4.3.1 Logistic Regression Analysis

Examination of each covariate with the response variable can provide a preliminary idea how important the variable is. Consequently, a univariate logistic regression model was fitted and variables with p-value < 0.25 were considered as candidates for the multiple logistic regression model (Hosmer and Lemeshow, 2000). Multiple logistic regressions is used when the response (*herdposit* / *abortion*) is dichotomous and the explanatory variables are of any type, qualitative, quantitative or both. It can be used not only to identify risk factors but also to predict the probability of success.

The general multiple logistic regression model is given as:

$$\log it[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Where, $\pi(x)$ is the probability of the herd is *seropositive* or the herd has *abortion*, x_i's are covariates and $\beta_i's$ are their respective parameters.

The *backward selection procedure* was used to build the model to identify the primary important risk factors without clustering effect. Eventually, variables with p-value < 0.05 were retained for further statistical analysis.

The traditional standard error estimates for logistic regression models based on maximum likelihood from independent observations is no longer appropriate for data sets with cluster structure since observations in the same clusters tend to have similar characteristics and are more likely correlated with each other. Since unions are the sampling units, observations on herds are not independent. Ignoring clustering in analyses may exaggerate the precision, so risk factors are reported as significant even when this may not be correct, and it may also affect point estimates (Bennett *et al.,* 1991; Faes *et al.,* 2006).

2.4.3.2 Generalized Estimating Equations (GEEs)

Marginal models (also known as population-averaged models) are models in which responses are modeled marginalized over all other responses; the association structure is then typically captured using a set of association parameters, such as correlations, odds ratios, etc (Molenberghs and Verbeke, 2005). So, the marginal model is used when the researcher investigates the population and wishes to model the population averaged response as a function of the covariates while accounting for the correlations in the data. It means that the existence of clustering is recognized but considered a nuisance characteristic. One such model is Generalized Estimating Equations (GEEs) (Liang and Zeger, 1986).

Liang and Zeger (1986) proposed the use of generalized estimating equations (GEEs) to analyze clustered binary data and are modeled with the same link function and linear predictors (systematic component) as in the independence case (logistic regression). The *logit* link function is

$$g(\mu_{ij}) = \boldsymbol{x_{ij}^T \beta}$$

Where, the regression parameters $\boldsymbol{\beta}$ are estimated by solving the estimating equations (or score equations) as

$$\sum_{i=1}^{N} \frac{d\mu_i}{d\beta} V_i^{-1}(y_i - \mu_i(\beta)) = 0$$

Where, $V_i = \boldsymbol{A_i^{1/2} R_i A_i^{1/2}}$ is marginal covariance matrix of y_i with $\boldsymbol{A_i}$ is the matrix of marginal variances (which are the same as for logistic regression) on the main diagonal and zeros elsewhere and $\boldsymbol{R_i}$ is the marginal correlation matrix. But the correlations in the data are specified by adopting a so-called working correlation assumption about the association structures. Typical correlation structures for clustered data are the independence, exchangeable and compound symmetry structure. In addition, GEE is a non-likelihood method that corrects for the clustering effect (correlation structure) and uses correlation to capture the association within the clusters. The parameter estimates β are consistent even if the working correlation matrix is misspecified but loss of efficiency in β may result from a poor or incorrect choice of the covariance structure (Liang and Zeger, 1986; Molenberghs and Verbeke, 2005). In this study, responses from herds in the same union are correlated and GEE could be used to fit a marginal model for factors that are associated with brucellosis and abortion.

2.4.3.3 Cluster weighted Generalized Estimating Equations (CWGEE)

Using GEE, the correlation between cluster members is modeled in order to determine the weight that should be assigned to the data from each cluster. If the outcome measured among cluster members is independent of cluster size (i.e., if cluster size is uninformative), clustering only enters the analysis to obtain a valid variance estimates (using the sandwich variance estimator). The inference is valid even if the working correlation is misspecified. However, if cluster size is informative (cluster size is related to the outcome of interest), then the different ways of weighing the data result in different marginal models. In that case, the choice of a working correlation matrix becomes an important issue and inappropriate choice resulting in misleading and biased parameter estimates (Williamson *et al.,* 2003; Aerts *et al.,* 2010).

Williamson *et al.* (2003) demonstrate that two marginal analyses can be of interest in the case of informative cluster size (*Union size*). Firstly, one might be interested in the probability of an arbitrary randomly sampled herd from the full population of herd. Secondly, interest can be in the probability of an arbitrary herd at random from a randomly selected union (first sample an arbitrary union, next given that union sample an arbitrary herd). So, the two approaches are: i) GEE with independence working correlation weighing each cluster (union) member equally, and ii) GEE with independence working correlation using weights inversely proportional to the cluster (union) size $1/n_i$. This approach is indicated as cluster weighted generalized estimating equation (CWGEE). These two marginal analyses will have the same asymptotic parameter estimates, except when cluster size is related to the outcome. Standard GEE approach provides parameter estimates that are weighted by the clusters and will no longer yield consistent estimates. When interested in the probability of an arbitrary randomly selected herd from a randomly selected union, the GEE method is no longer valid. Therefore, by incorporating the cluster size as a covariate in the model, the GEE method will again yield unbiased estimates of β (Faes *et al.,* 2006; Aerts *et al.,* 2010). Moreover, when cluster size is uninformative, unweighted or CWGEE analyses produce equivalent results, and the GEE analyses may be optimized by using a more appropriate working correlation than the one corresponding to independence.

2.4.3.4 Generalized Linear Mixed Model (GLMM)

Herds belonging to a union (cluster) share the same environment (grazing places) as well as characteristics such as the type of farm (subsistence or commercial) and other unobserved factors (Speybroeck *et al.,* 2003). A random-effect or cluster-specific model describes the dependencies between responses because of shared factors in a union. So, this paper also focuses on cluster-specific approaches and within-cluster covariate effects.

The Generalized Linear Mixed Model (GLMM) has become a popular approach to modeling correlated discrete data. The GLMM can be seen as an extension of the generalized linear model and account for correlation among clustered observations by adding random effects to the linear predictors. In a random effects model, it is assumed that there is natural heterogeneity across the clusters. This heterogeneity can be modeled by a probability distribution which implies that the regression coefficients are varying from one cluster to another. Conditionally on random effects for each cluster, it assumes that the responses across the cluster are independently distributed as

$$Y_{ij}|\boldsymbol{b_i} \sim Bernoulli\ (\pi_{ij}),$$

$$\eta_{ij} = log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = X_{ij}\beta + Z_{ij}\boldsymbol{b}_i,$$

Where Y_{ij} is the *j-th* outcome observed for cluster (subject) *i, i =1,...,N,* $j = 1,...,n_i$. The model is completed by assuming that, conditionally on the subject-specific effects $\boldsymbol{b}_i$, a random vector which is assumed to be normally distributed with mean vector **0** and covariance matrix **D**, the responses Y_{ij} are independent. X_{ij} and Z_{ij} are $(n_i \times p)$ and $(n_i \times q)$dimensional vectors of known covariates. Similarly, β is a *p*-dimensional vector of unknown fixed effect regression parameters (Molenberghs and Verbeke, 2005).

Random effects models can be fitted by maximization of the likelihood obtained by integrating out the random effects using numerical approximations. The `PROC NLMIXED` procedure with Adaptive Gaussian Quadrature method is used for fitting the GLMM approach. In the model building process, the backward selection procedure is used and compare with AIC ($AIC = 2k - 2lnL, k\ is\ the\ no.of\ parameters\ and\ L\ is\ the\ likelihood\ function$) values. The best model is the one with minimum AIC value.

2.4.3.5 Joint Modeling of two binary outcomes: Random effect approach

The main objective of the joint modeling is to provide a framework within which questions of scientific interest pertaining to systematic relationships among the multiple outcomes and between them and other factors (*breeding*, *herdsize* etc.) may be formalized. To ensure valid inferences, joint models must appropriately account for the correlation among the outcomes (Fitzmaurice *et al.,* 2008). In this study the interest is of the association between the outcomes, brucellosis seropositivity (*herdposit*) and *abortion* and also whether there is effect of the risk factors on the outcomes (*herdposit* and *abortion*).

According to Fieuws and Verbeke (2006), the joint mixed (generalized linear mixed) model assumes a mixed (generalized linear mixed) model for each outcome, and these 'univariate' models are combined through specification of a joint multivariate distribution for all random effects. Obviously, the joint model can be considered as a new mixed (generalized linear mixed) model.

Let $Y_{i1j} = 1$ if the j-th herd in the i-th union is seropositive and $Y_{i1j} = 0$ otherwise; and let $Y_{i2j} = 1$ if the j-th herd in the i-th union is reported to have abortion and $Y_{i2j} = 0$ otherwise. We denote the sequence for cluster *i* by $Y_{i1} = \left(Y_{i11}, Y_{i12}, \dots, Y_{i1n_i}\right)'$ and $Y_{i2} = \left(Y_{i21}, Y_{i22}, \dots, Y_{i2n_i}\right)'$. Random-effects models are the most frequently used models to analyze longitudinal and clustered data. A mixed model also allows incorporation of different types of outcomes of different nature in a uniform and natural way (Molenberghs and Verbeke, 2005). Let us first define the model in the general setting of two outcomes. For the bivariate response vector $Y_i = (Y_{i1}', Y_{i2}')'$ we can assume a generalized mixed model of the form

$$Y_i = \mu_i(\eta_i) + \varepsilon_i = h(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i) + \varepsilon_i \dots\dots\dots (1)$$

where μ_i's are the fixed and random effects and ε_i is the residual error structure. The components of the residual error structure ε_i have the appropriate distribution with variance depending on the mean-variance relationship of the various outcomes, and can contain in a correlation matrix R_i and an overdispersion parameter. The components of the inverse link function $h(.)$ depend on the nature of the outcomes in Y_i. For example, in our case two responses are binary and the link functions are the logit link for both. X_i and Z_i are $(2n_i \times p)$ and $(2n_i \times$

q) dimensional matrices of known covariate values corresponding to subject i, and β a p-dimensional vector of unknown fixed regression coefficients. Furthermore, $b_i \sim N(0; \boldsymbol{D})$ are the q-dimensional random effects. Since our interest is in the correlation structure of the data, a general first-order approximate expression for the variance-covariance matrix of $\boldsymbol{Y}_i$ is derived

$$V_i = Var(\boldsymbol{Y}_i) \simeq \Delta_i Z_i D Z_i' \Delta_i' + \Sigma_i \dots\dots\dots (2)$$

with $\Delta_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right) | b_i = 0$ and $\Sigma_i \simeq \Xi_{\mathrm{i}}^{1/2} A_i^{1/2} R_i(\alpha) A_i^{1/2} \Xi_{\mathrm{i}}^{1/2}$ where A_i is a diagonal matrix containing the variances following from the generalized linear model specification of Y_{ik} (k = 1, 2), given the random effects b_i = 0, i.e., with diagonal elements $v(\mu_{ik}|b_i = 0)$. Likewise, Ξ_{i} is a diagonal matrix with the overdispersion parameters along the diagonal. The 1st term at the right hand side of (2) corresponds to the random-effects structure $h(X_i\beta + Z_i b_i)$; the second term at the right hand side of (2) captures the variance-covariances in the residual error ε_i. From equation (2) it is clear that the correlation between the outcomes can be modeled either using the residual variance of Y_i or through specification of the random-effects structure $Z_i b_i$. When there are no random effects in (1), a marginal model is obtained. When there are no residual correlations in R_i , this is called conditional independence model or purely random-effects model which is denoted by GLMM (Molenberghs and Verbeke, 2005; Faes *et al.,* 2008; Fitzmaurice *et al.,* 2008).

More specifically, we formulate a possible joint model for two binary outcomes, while accounting for the clustering nature of the outcomes, using a conditional independence random-intercepts model with a general variance-covariance matrix $\boldsymbol{D}$ and residual correlation matrix$R_i(\alpha) = I$.

Therefore, a GLMM can be assumed with correlated random effects as

$$\begin{pmatrix} Y_{i1j} \\ Y_{i2j} \end{pmatrix} = \begin{pmatrix} \dfrac{\exp(\alpha_0 + \alpha_1 X_{ij} + b_{i1})}{1 + \exp(\alpha_0 + \alpha_1 X_{ij} + b_{i1})} \\ \dfrac{\exp(\beta_0 + \beta_1 X_{ij} + b_{i2})}{1 + \exp(\beta_0 + \beta_1 X_{ij} + b_{i2})} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \dots\dots\dots (3)$$

Where the random effects b_{i1} and b_{i2} are normally distributed as

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right\}$$

and where ε_{i1} and ε_{i2} are independent. It is assumed that

$$Var(\varepsilon_{i1j}) = v_{i1j} = \pi_{i1j}(b_{i1} = 0)[1 - \pi_{i1j}(b_{i1} = 0)] \text{ and}$$

$$Var(\varepsilon_{i2j}) = v_{i2j} = \pi_{i2j}(b_{i2} = 0)[1 - \pi_{i2j}(b_{i2} = 0)]$$

The variances of Y_{i1j} and Y_{i2j} can be calculated from equation (2) in which

$$Z_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Delta_{\text{ij}} = \text{A}_{\text{ij}} = \begin{pmatrix} v_{i1j} & 0 \\ 0 & v_{i2j} \end{pmatrix}, D = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}, \ \Xi_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and the approximate variance-covariance matrix is

$$V_{ij} = \begin{pmatrix} v_{i1j}^2\tau_1^2 + v_{i1j} & \rho\tau_1\tau_2 v_{i1j}v_{i2j} \\ \rho\tau_1\tau_2 v_{i1j}v_{i2j} & v_{i2j}^2\tau_2^2 + v_{i2j} \end{pmatrix}$$

and the approximate correlation between the two outcomes is

$$\rho_{Y_1Y_2} = \frac{\rho\tau_1\tau_2 v_{i1j}v_{i2j}}{\sqrt{v_{i1j}^2\tau_1^2 + v_{i1j}}\sqrt{v_{i2j}^2\tau_2^2 + v_{i2j}}}$$

In the case of conditional independence ($\rho \equiv 0$), the approximate marginal correlation function $\rho_{Y_1Y_2}$ also equals zero. In the case of $\rho \equiv 1$, this model reduces to a shared-parameter model with scale factor λ is equal to τ_1/τ_2. SAS procedure `PROC NLMIXED` was used to obtain parameter estimates for this bivariate model (Faes *et al.,* 2008; Fitzmaurice *et al.,* 2008).

2.5 Software

The well known statistical packages SAS (version 9.2) and STATA (version 9) were used to analyze the data. 5% level of significance was also used throughout the study.

3. Results

3.1 Exploratory Data Analysis (EDA)

Of the 388 herds investigated, 35.10% were brucellosis seropositive and 20.62% herds reported abortions in the study area of *Mymensingh* and *Sherpur* districts. Figure 1 shows the distribution of seroprevalence of brucellosis seropositive according to the study areas (Unions). It can be seen in the graph, that there were some variations of seroprevalence of brucellosis among the unions. Highest seroprevalence of brucellosis (87.50%) was observed in *Bhabkhali* union followed by *Dakatia* (78.30%), *Noyabil* (76%) and *Rambhadrapur* (71.40%) unions.

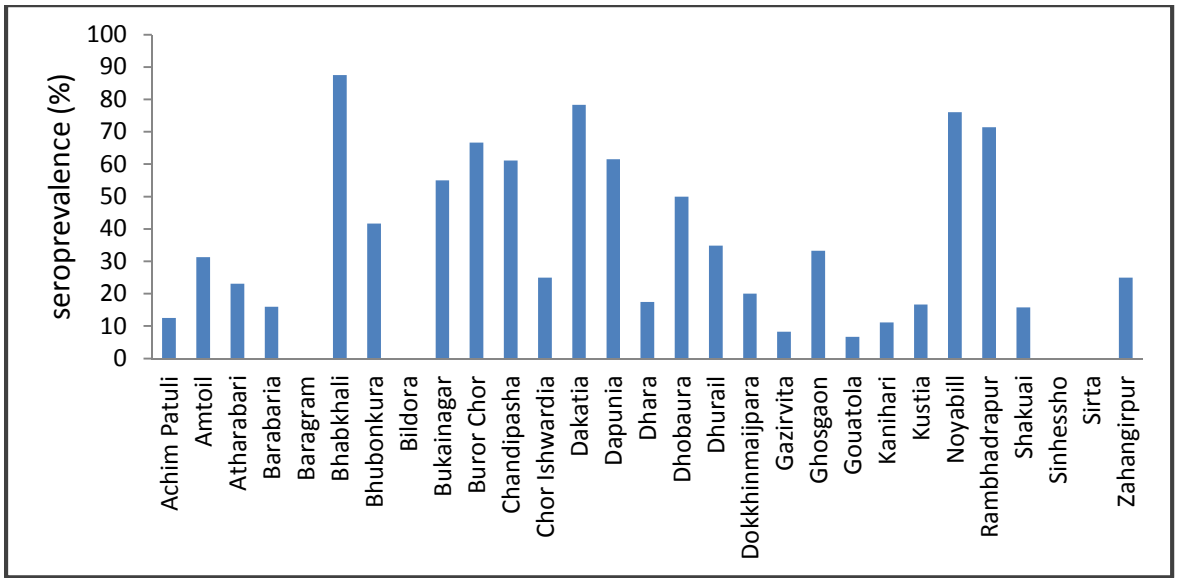


Figure 1 : Seroprevalence of brucellosis seropositivity according to union

Figure 2, on the otherhand, shows the prevalence of herds reporting abortions among the unions. *Baragram* union investigated only one herd and it had the history of abortion. The next higher prevalence of reporting abortions was *Noyabil* (48.0%) followed by *Bukainagar* (45.0%), *Kanihari* (40.74%) unions. It was noted that history of herd' abortion differed among the unions.

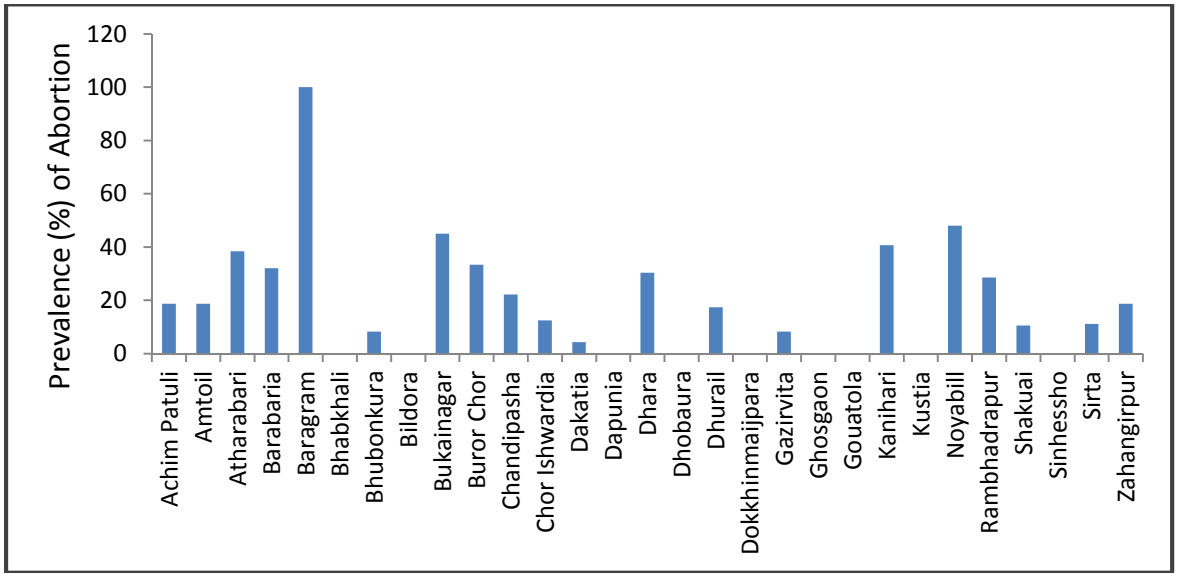


Figure 2: Prevalence of abortion according to union

The herds which were artificially inseminated (AI) had a higher (37.3%) chance of being positive for brucellosis as compared to that of natural services (NS) (*Figure 3a*). Similarly, the prevalence of abortion is 40.7% in the herds which used AI and only 8% for those which used

their own bulls (natural services) for breeding (*Figure 3b*). If the farm owner purchased animals in their farm, then the whole farm is treated as purchased (purchase = yes) otherwise not. Purchasing of animals is an important risk factor for brucellosis seropositivity. Figure 3c shows that the herd with purchased animals reported 32.1% of brucellosis seropositivity whereas those that did not purchase had 36.2% of brucellosis seropositivity. In the case of abortion, the prevalence is higher (29.5%) for herds which purchased animals than for those that did not purchase animals (*Figure 3d*).

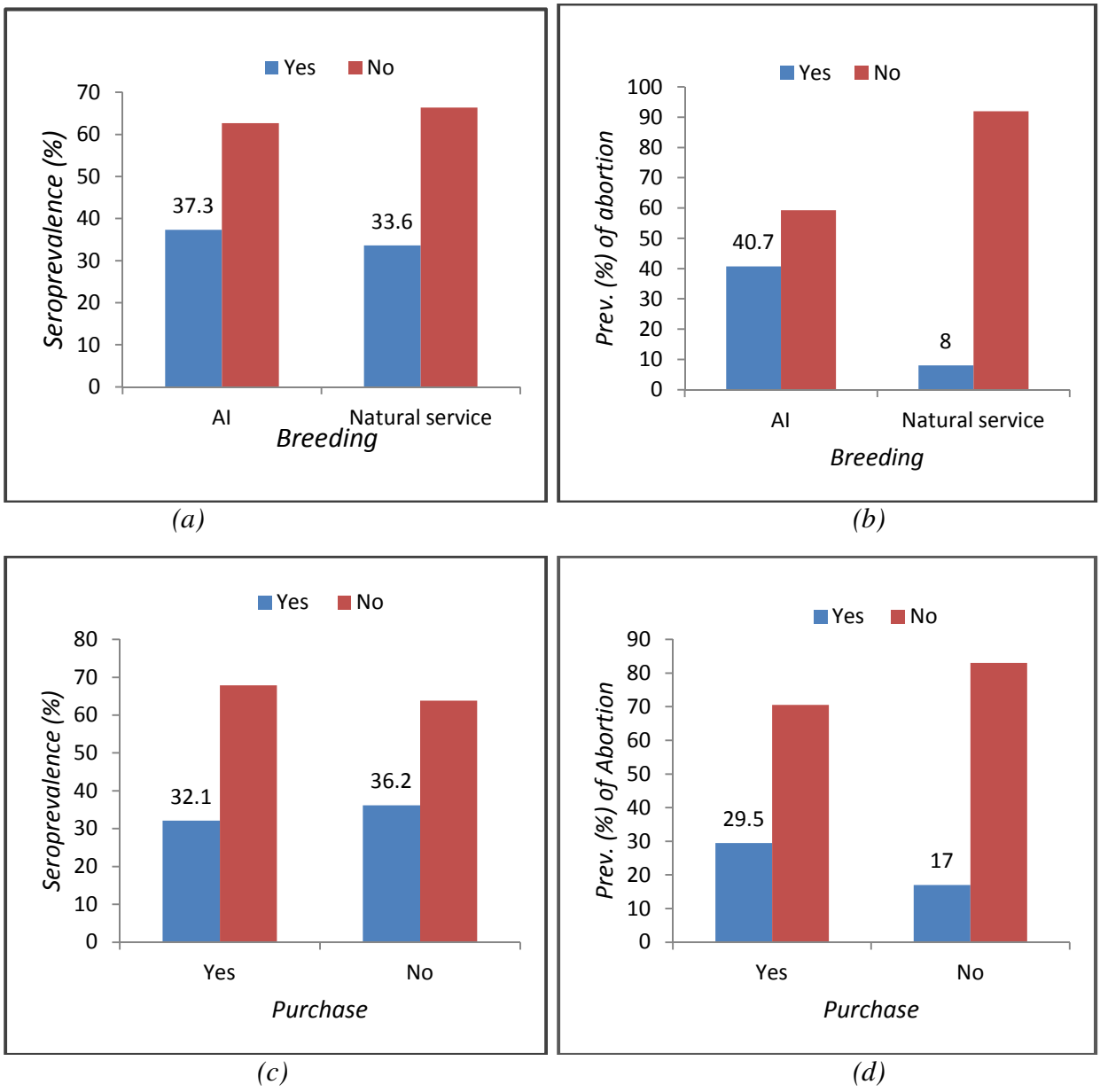


Figure 3: Herd prevalence of brucellosis and abortion with (a)-(b) Breeding and (c)-(d) Purchase of animals respectively

The *Brucella* seroprevalence of subsistence farm was 31.1% whereas the seroprevalence for commercial farm was 65.9%. It was observed that commercial farms consisted of a large number of cows. As there were more cows, the chance of being infected of that herd increases. But in case of abortion, the subsistence farms had reported 21.2% abortions while the commercial farm had 15.9% reported abortions.

The minimum and maximum *herdsize* (no. of cattle) was 1 and 13 respectively. The variable *herd size* was categorized into two groups according to their average. For the herd that consisted of more than 4 cows, the chance of being brucellosis seropositive (49.3%) was higher than the herd made of 3 or less number of cows. The average union size (no. of herds) was 1989.69 with their standard deviation 572.9. This variable also categorized into three groups according to quartiles. But both variables were used as continuous in the statistical analysis. The herds of the selected union make up the cluster and the observation within the cluster is a single herd. Therefore, the maximum cluster size (union size) was 3000 herds. The prevalence of brucellosis seropositivity was higher in *medium* union size (41.4%) than *small* (28.0%) and *large* union size 33.5% (*Table2*).

In the case of *abortion*, herd prevalence of abortion was higher (26.0%) in herd size *≤3* than that of herd size *≥4*. The prevalence of reported abortion increases when the number of herds per union increases, indicating that the cluster size (*unionsize*) may be non-ignorable for this study (*Table2*).

Table 2: Distribution of herd prevalence of brucellosis seropositivity and abortion

Variables	Seropositivity		Prevalence	Abortion		Prevalence
	No	Yes	(%)	No	Yes	(%)
Herd Size (no. of cattle)						
≤3	178	64	26.4	179	63	26.0
≥4	74	72	49.3	129	17	11.6
Mean ± SD	3.46 ± 2.33					
Union Size (no. of herds)						
Small (700-1400)	59	23	28.0	72	10	12.2
Medium (1401-2000)	78	55	41.4	109	24	18.0
Large (2000+)	115	58	33.5	127	46	26.6
Mean ± SD	1989.69 ± 572.9					

3.2 Statistical Analysis

3.2.1 Logistic Regression Analysis

Since the outcome variable *herdposit* is binary, a logistic regression model is initially carried out. In the first step, an ordinary logistic regression model is performed, ignoring the fact that herds can come from the same union. In order to have a valid model, variables to be included in the model have to be appropriate and moderate in number. Thus, variable reduction process was performed by fitting univariate logistic regression for each covariate and variables with p-value > 0.25 were dropped. Only the variable *purchase* (p-value=0.4446) was dropped in this procedure. The variable *breeding* was also not significant but biologically it has a great impact on brucellosis seropositivity. In that sense, *breeding* was included in the multiple logistic regression model. Finally, *breeding, farmtype, herdsize, unionsize* and their two way interactions were used in multiple logistic regression. Using manual *backward selection procedure,* variables with p-value < 0.05 were retained for further statistical analysis. Therefore, the final multiple logistic regression model is

$$logit(herdposit) = \beta_0 + \beta_1 * breeding + \beta_2 * farmtype + \beta_3 * herdsize + \beta_4 * unionsize + \beta_5 * unionsize * farmtype + \beta_6 * breeding * herdsize$$

Fitting this model leads to the results in Table 3. Here, parameters estimates are given, together with their standard errors ignoring clustering in the data. The parameter estimates of the interaction terms (*farmtype* and *unionsize*) and (*breeding* and *herdsize*) are significantly different from zero. The standard errors of parameter estimates are not consistent, i.e. it underestimates the standard errors (by heterogeneity factor, Pearson chi-square/DF, $\varphi = 1.213 > 0$ indicating overdispersion) when ignoring clustering in the data but the point estimates of the parameters may remain consistent.

Table 3: Parameter estimates and standard errors (for herdposit)

Parameter	Estimate	Std. Error	Pr > ChiSq
Intercept	0.7182	1.3371	0.5912
Breeding	-0.6338	0.4477	0.1569
Farmtype	-2.9059	1.3208	0.0278
Herdsize	0.1226	0.0814	0.1321
Unionsize	-0.0008	0.0006	0.1833
*Farmtype*unionsize*	0.0013	0.0006	0.0423
*Breeding*herdsize*	0.2571	0.1135	0.0235

Similar procedure was used in the case of *abortion* status of herds as response. So, the final multiple logistic regression model is

$$logit(abortion) = \beta_0 + \beta_1 * breeding + \beta_2 * purchase + \beta_3 * herdsize + \beta_4 * unionsize + \beta_5 * breeding * herdsize$$

Table 4 gives the parameter estimates and their standard errors ignoring the clustering effect. *Breeding, purchase, union size* and the interaction between *breeding and herdsize* are primarily important risk factors for reported *abortion* in the herds.

Table 4: Parameter estimates and their standard errors (for abortion)

Parameter	Estimate	Std. Error	Pr > ChiSq
Intercept	-4.0064	0.7502	<.0001
Breeding	3.3136	0.5650	<.0001
Purchase	0.5939	0.3075	0.0534
Herdsize	0.0440	0.1005	0.6616
Unionsize	0.0006	0.0003	0.0306
Breeding herdsize*	-0.4074	0.1453	0.0050

Results from Table 4 show that *unionsize,* the interaction between *breeding* and *herdsize* and *purchase* are significant factors for herd's reported abortion. But ignoring the clustered nature in the data, overestimates precision ($\varphi = 1.697 > 0$) and hence underestimates standard errors. Next we proceed with the methods which deal with clustering e.g. marginal model and random effect models.

3.2.2 Generalized Estimating Equations (GEEs)

Before going to the marginal model like GEE, first we have to check if the cluster size (union size) is informative or not. The following GEE model with independence working correlation assumption was fitted to the binary variable *herdposit:*

$$P(herdposit = 1|breeding, farmtype, herdsize, union\ size) = expit(\beta_0 + \beta_1 * breeding + \beta_2 * farmtype + \beta_3 * herdsize + \beta_4 * unionsize + \beta_5 * farmtype * unionsize + \beta_6 * breeding * herdsize)$$

Table 5: Parameter estimates for checking informative cluster size (union size) on seropositivity

Parameter	Estimate	Standard Error	Pr > |Z|
Intercept	0.7182	1.4710	0.6254
Breeding	-0.6338	0.3719	0.0883
Farmtype	-2.9059	1.4660	0.0475
Herdsize	0.1226	0.0804	0.1271
Unionsize	-0.0008	0.0007	0.2914
*Unionsize*farmtype*	0.0013	0.0007	0.0857
*Breeding *herdsize*	0.2571	0.1094	0.0188

The model was fitted based on the primary important risk factors associated with brucellosis seropositivity. Table 5 gives the result of the parameter estimates with their standard errors. The cluster size (*unionsize*) or even the interaction effects with *farmtype* are not significantly related with brucellosis seropositivity indicating that the cluster size (*unionsize*) is uninformative. Now, we proceed on the unweighted GEE to identify the risk factors associated with brucellosis.

Since the *unionsize* is uninformative, the goal is to recognize the association between covariates and randomly selected herd from the overall population herds, a GEE that uses the independence working correlation will provide the valid result. In order to build the unweighted GEE model, the primary risk factors and their two-way interactions were included in the model. The manual backward selection procedure was used to identify the important risk factors associated with brucellosis seropositivity. The final model gives the empirical based parameter estimates and their standard errors noted in Table 6. *Herdsize* and interaction effect with *breeding* are significant risk factors with brucellosis seropositivity. Moreover, interaction between *herdsize* and *breeding* has positive effect on brucellosis.

Table 6: Parameter estimates (standard error) according to different GEE methods for brucellosis seropositivity

Parameter	*Unweighted GEE*		*Cluster weighted GEE*	
	Estimate (S.E)	P-value	Estimate (S.E)	P-value
Intercept	-1.3253 (0.3724)	0.0004	-1.4924(0.4309)	0.0005
Breeding	-0.5210 (0.4032)	0.1962	-0.3954 (0.3893)	0.3098
Herdsize	0.1740 (0.0679)	0.0104	0.1957 (0.0785)	0.0127
*Herdsize*breeding*	0.2280 (0.1095)	0.0373	0.2207 (0.1025)	0.0313

Similarly, to check if the cluster size informative or not in the case of response *abortion*, the GEE method with independent working correlation assumption was fitted as

$$P(abortion = 1|breeding, purchase, herdsize, union\ size) = expit(\beta_0 + \beta_1 * breeding + \beta_2 * purchase + \beta_3 * herdsize + \beta_4 * unionsize + \beta_5 * breeding * herdsize)$$

Table 7: Parameter estimates for checking informative cluster size (union size) on Abortion

Parameter	Estimate	Standard Error	Pr > |Z|
Intercept	-4.0064	0.9714	<.0001
Breeding	3.3136	0.8631	0.0001
Purchase	0.5939	0.3538	0.0932
Herdsize	0.0440	0.1094	0.6875
Unionsize	0.0006	0.0003	0.0219
*Herdsize*breeding*	-0.4074	0.2154	0.0585

It can be seen from Table 7 that the significance of *unionsize* as a main effect indicates that the cluster size (*unionsize*) is non-ignorable i.e. the *unionsize* is related with the reported *abortion* in herd. So, the conclusion is that there is an overall positive effect of *unionsize* i.e. the larger the *unionsize*, the larger the probability for a herd of that union to have a previous record of abortion. In that case two marginal analyses can be of interest.

3.2.3 Cluster-weighted Generalized Estimating Equations (CWGEE)

Although the cluster size (*unionsize*) was independent of brucellosis seropositivity, we fitted cluster-weighted GEE to compare the parameter estimates. It can be seen in Table 6 that all the parameter estimates in both unweighted GEE and CWGEE are similar in their magnitude.

In the case of reported abortion in the herd, we are interested in two explanations due to informative cluster size. One might be interested in the probability of a randomly selected herd from the total population herds. Another is the probability of a randomly selected herd from a randomly selected union. In this study, there is a positive association between the size of the union and the prevalence of reported abortion among the herds in that union. So, GEE method with independence working correlation might overweigh individuals in larger clusters, resulting in an overfitting of the prevalence of abortion (Faes *et al*., 2006; Williamson *et al.,* 2003). To solve this issue, cluster weighted GEE, proposed by Williamson *et al.* (2003) was fitted where

the contribution to the estimating equation from a union is weighted by the inverse of the cluster size. In this way, all unions are given equal weights and individuals in large clusters are no longer overweighed. Table 8 provides the parameter estimates and their standard errors for both unweighted GEE and CWGEE. CWGEE is considered only when *unionsize* is not included as a covariate in the model. But we are more interested to incorporate the cluster size (*unionsize*) as a covariate in the model. In that case, the probability of randomly selected herd from the total population gives more specific result and therefore unweighted GEE method will again provide unbiased parameter estimates. So, an unweighted GEE with independence working correlation assumption was fitted with inclusion of *unionsize* as a covariate and the results are presented in Table 8. *Breeding, unionsize* and interaction effect of *breeding* and *herdsize* were significantly associated with the prevalence of reported abortion in the herds.

Table 8: Parameter estimates (standard error) according to different GEE methods for Abortion

Parameter	*Unweighted GEE*		*Cluster weighted GEE*	
	Estimate (S.E)	P-value	Estimate (S.E)	P-value
Intercept	-3.8611 (0.9409)	<.0001	-2.5319 (0.5358)	<.0001
Breeding	3.4652 (0.8759)	<.0001	3.2466 (0.8008)	<.0001
Herdsize	0.0649 (0.1137)	0.5681	0.0286 (0.1001)	0.7753
Unionsize	0.0006 (0.0003)	0.0248	…	…
*Herdsize*breeding*	-0.4538 (0.2238)	0.0425	-0.4608 (0.2291)	0.0443

3.2.4 Generalized Linear Mixed Model (GLMM)

So far, we have dealt with marginal models to identify the risk factors associated with brucellosis and abortion in the herd. However, it is important to investigate which risk factors are related to the prevalence of brucellosis and abortion within the specific union and to explain the differences between unions. The random effects model, specifically, the random intercept model was considered to account for heterogeneity among unions and to make union specific inferences. In the model building process, we incorporated all the main and two-way interaction effects and used backward selection procedure and compare with AIC values to come up with the final model. The saturated model provided the AIC value equal to 435.6.

The final fitted GLMM model for brucellosis seropositivity is given by

$$Y_{ij}|b_i \sim Bernoulli(\pi_{ij})$$

$$logit(\pi_{ij}) = \beta_0 + \beta_1 * breeding + \beta_2 * herdsize + \beta_3 * breeding * hersize + b_i$$

This model gives the smallest AIC value equal to 433.7. The parameter estimates and their standard errors are presented in Table 9. The result shows that only *herdsize* is significantly related to the seroprevalence of brucellosis and it has also a positive effect of being infected with brucellosis. Though *breeding* is an important factor for the disease brucellosis, it is however not significant here. The intra-union correlation coefficient ($\rho = \frac{\sigma_b^2}{\sigma_b^2 + \pi^2/3}$), estimated from the random-intercepts model equals 0.2693, indicating that herds within unions are correlated and is highly significant (*Table 10*).

In the case of reported abortion, *breeding* and interaction effect with *herdsize* are highly significant with the prevalence of abortion in the herd. The interaction effect has a negative effect on abortion whereas the main effect *breeding* has a positive effect on reported abortion (*Table 9*). This model is the best fitted model with smallest AIC value equal to 318.9 as compared with the saturated model with AIC value equal to 320.3. The intra-cluster correlation coefficient (0.1037) indicates that the herds within the unions are also slightly correlated but not significant (*Table 10*).

Table 9: Parameter estimates and their standard errors for GLMM

Parameter	*Herd seropositivity*			*Abortion*		
	Estimate	Std. Error	P-value	Estimate	Std. Error	P-value
Intercept	-1.5815	0.3889	0.0004	-2.9108	0.4913	<.0001
Breeding	-0.6478	0.5228	0.2257	3.5995	0.5893	<.0001
Herdsize	0.2245	0.0719	0.0041	0.0688	0.0998	0.4966
*Breeding*herdsize*	0.1686	0.1256	0.1901	-0.4820	0.1517	0.0036
Sigma	1.1012	0.2264		0.6170	0.2214	
σ_b^2	1.2127	0.4985	<0.0001	0.3806	0.2732	<0.0001

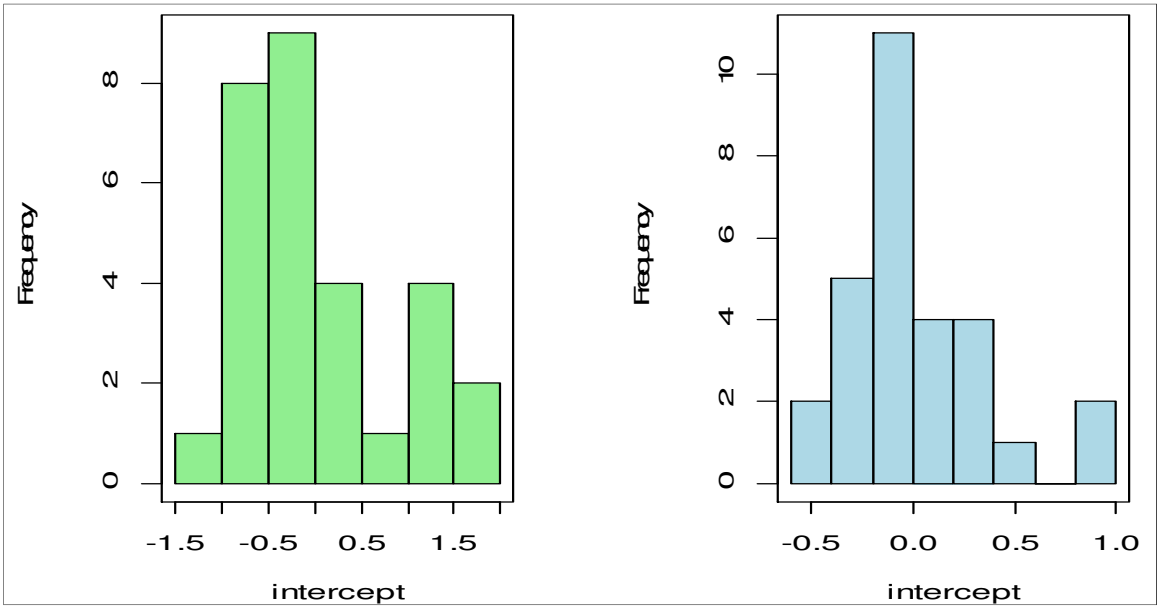
Table 10: Intra-cluster correlation coefficient

Parameter	*Herd seropositivity*			*Abortion*		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
ICC (rho)	0.2693	0.0808	<0.0025	0.1037	0.0667	0.1314

In addition, the variability around the random intercept is found to be significant which could be regarded as evidence for taking the clustering into account. It is worth noting that the p-value

associated with random intercept variance, σ_b^2 in Table 9 is based on mixture of chi-squares $(0.5 * \chi_0^2 + 0.5 * \chi_1^2)$ since the null hypothesis of zero variance lies on the boundary of the parameter space, making the original p-value conservative.

Figure 4 presents the histograms of random intercepts for unions. Empirical Bayes estimates of the random intercepts for both responses do not show presence of outlying union. But both somehow shows a little bit of skewness to the right.



(a) *(b)*

Figure 4: Histogram of random intercepts for (a) seropositivity and (b) abortion

3.2.5 Joint Modeling of two binary outcomes: Random-effects model

In order to know the correlation between reported abortion and brucellosis seropositivity in the herd and whether there is significant effect of the risk factors on the two outcomes, the joint random effects model of two binary responses was fitted. We fitted separate analysis at each outcome in *Section 3.2.4*. But this was not an efficient method and might not capture all effects present in the dataset. Because of this, we focus on the joint random intercepts (union-specific random intercepts) model accounting for possible association between *abortion* and brucellosis seropositivity (*herdposit*) in the herd.

Table 11: Parameter estimates for joint model of two binary responses

Parameter	Estimate	Standard Error	Pr > |t|
	Abortion		
Intercept	-2.9089	0.4915	<.0001
Breeding	3.5980	0.5894	<.0001
Herdsize	0.0688	0.0998	0.4963
*Breeding*herdsize*	-0.4831	0.1523	0.0038
Var (τ_1)	0.3790	0.2731	
	Herdposit		
Intercept	-1.5817	0.3890	0.0004
Breeding	-0.6503	0.5237	0.2250
Herdsize	0.2244	0.0719	0.0042
*Breeding*herdsize*	0.1692	0.1258	0.1897
Var (τ_2)	1.2131	0.4988	
	Association		
Cov (τ_1, τ_2)	0.0223	0.2579	
Correlation (ρ)	0.0219	0.2535	0.9318

Table 11 shows the parameter estimates from the joint random intercepts model. The correlation among the two responses is 0.0219 indicating negligible association and is not statistically significant (p-value = 0.9318). The parameter estimates are similar to that of univariate cases, with the interaction between *breeding* and *herdsize* significantly associated with reported abortion in the herds. Similarly, only *herdsize* is significantly associated with brucellosis seropositivity.

Plotting the two random intercepts from Empirical Bayes estimates also shows that there is no association between brucellosis seropositivity and abortion (*Figure 5*).

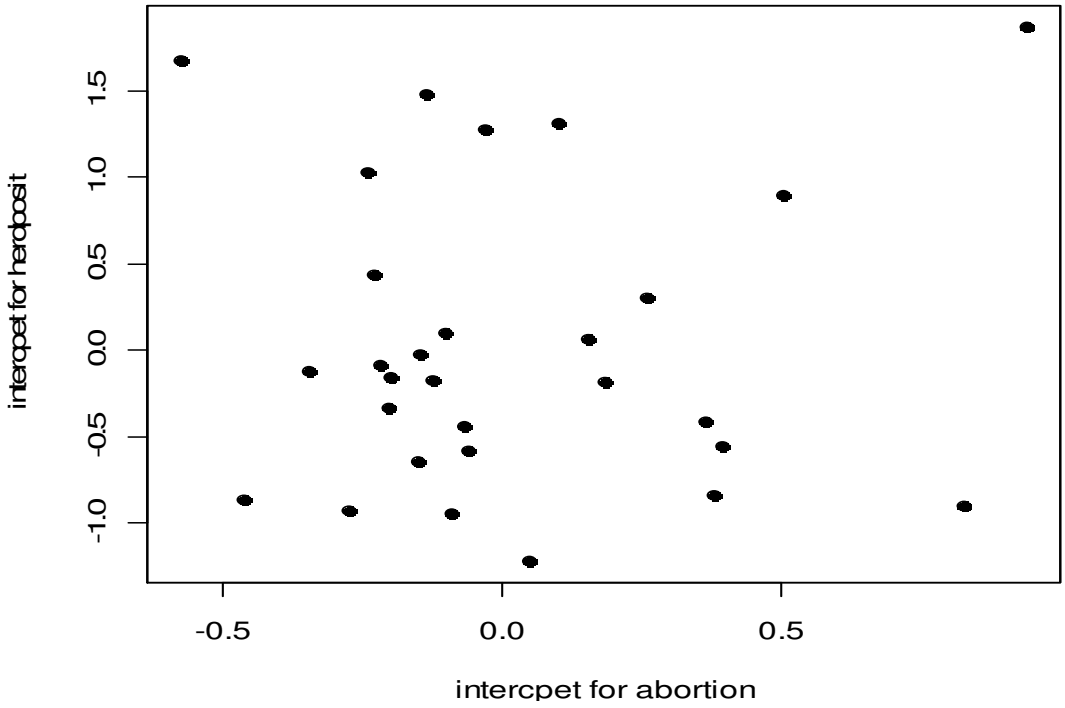


Figure 5: Scatter plot of Empirical Bayes estimate for two random intercepts

4. Discussion, Conclusion and Recommendation

4.1 Discussion

The objective of this report was to identify the herd level risk factors associated with bovine brucellosis seropositivity and abortion in Bangladesh. A cross-sectional sero-epidemiological study of bovine brucellosis was conducted between September 2007 and August 2008 in the *Mymensingh* and *Sherpur* districts of Bangladesh. A total of 388 herds were selected within the 29 unions. Serum samples were collected from all the cows of 388 herds and tested for *Brucella* infection by the Rose Bengal Plate Test (RBT), a Serum Agglutination Test (SAT) and an indirect Enzyme-linked Immunosorbent Assay (iELISA) respectively. A seropositive herd was defined as one in which at least one animal tested positive within the herd on either three tests (RBT, SAT or iELISA). While a herd was defined as positive for abortion if at least one cow within the herd was reported to have a previous record of abortion.

Both descriptive statistics and graphical representations were employed to explore the data. The exploratory analysis showed that the overall herd level prevalence of bovine brucellosis and abortion were very high (35.10% and 20.62%). *Brucella* seropositive herds were present in

almost all the unions although some unions had high seroprevalence. A higher seroprevalence was observed in a large *herdsize* than in the small *herdsize,* which was consistent with several studies (Omer *et al.,* 2000; Matope *et al*., 2010). In the case of abortion, the prevalence was higher in small herd size. The prevalence of brucellosis seropositivity was also higher in *medium* union size than the *small* and *large* union size. Furthermore, the prevalence of reported abortion increased when the number of herds per union increased, indicating that the cluster size (*unionsize*) might be non-ignorable, which was later confirmed by statistical analysis. *Breeding* was also considered an important risk factor for both brucellosis and abortion. Additionally, the herds which used artificial insemination (AI) had a higher chance of being seropositive than those that made use of natural services (NS). Similarly, the herd prevalence of abortion was higher (40.7%) in herds that used AI than in those herds, which used their own bulls (natural services) for breeding. To confirm all these initial observations, proper testing, using appropriate models were done.

Correlated data are common in veterinary epidemiology, where clustered and hierarchical data are frequently observed. The study of bovine brucellosis was complicated because of clustering nature (herds within the union) in the dataset. Several techniques were employed to deal with clustering. The choice of analysis depends on the scientific goals; the most important techniques are population-averaged and random effect models.

Ignoring the clustering effect of union, a multiple logistic regression model was fitted to get the initial risk factors both for brucellosis seropositivity and abortion. But we may not interpret the results because of the standard errors of the parameter estimates were not consistent even though the point estimates were consistent. To solve this problem, first we considered the population-averaged method to identify the risk factors for both outcomes. Since our data was clustered by nature, it was necessary to check for informative cluster sizes (*unionsize*). Cluster size is informative when the response among observations in a cluster is associated with the cluster size (Williamson *et al*., 2003). To check the informativeness of the cluster size for both responses, the Generalized Estimating Equation (GEE) was fitted with the potential risk factors. The results showed that in the case of brucellosis seropositivity, the cluster size (*unionsize*) was uninformative but in the case of abortion, the *unionsize* was informative. In this study, both

unweighted and cluster weighted (CW) GEEs were fitted to identify the risk factors associated with bovine brucellosis and abortion. When cluster size is informative, a standard GEE will provide parameter estimates that are weighted by inverse of cluster size. If cluster size is uninformative then both unweighted and cluster weighted GEEs give same results otherwise they provide different results.

In the case of brucellosis seropositivity, *herdsize* and interaction effect with *breeding* were significant risk factors with brucellosis seropositivity. Moreover, interaction between *herdsize* and *breeding* has positive effect on brucellosis in both unweighted and CWGEE. The parameter associated with the interaction term indicates that for each unit increase in herd size and artificially inseminated herd, on average, the odds of having brucellosis seropositivity of herds is [*exp* (0.2280)] = 1.26 times higher than that of the odds of herd used natural service breeding (*Table 6*). Since union size is uninformative, the use of CWGEE will result in loss of efficiency but the interpretations are same as standard GEE.

In the case of herd reported abortion, we were interested in two interpretations due to informative cluster size. One might be interested in the probability of a randomly selected herd from the total population herds. Other is the probability of a randomly selected herd from a randomly selected union. For the second interpretation, cluster weighted GEE was fitted where the contribution to the estimating equation from a union is weighted by the inverse of the cluster size. In that case cluster size (union size) was not included in the model. In this way, all unions are given equal weights and individuals in large clusters were no longer overweighted. *Breeding* and interaction with *herdsize* were significant risk factors for abortion in CWGEE. But we were more interested to incorporate the cluster size (*unionsize*) as a covariate in the model. In that case, the probability of randomly selected herd from the total population gives more specific result based on the effect of union size on the reported abortion. So, an unweighted GEE with independence working correlation assumption was fitted with inclusion of *unionsize* as a covariate. The results indicated that *breeding, unionsize* and interaction effect of *breeding* and *herdsize* were significantly associated with the herd prevalence of reported abortion. There was an overall positive effect of union size on the herd reported abortion. It means that the larger the union, the larger the probability for a herd of that union to have a previous record of abortion. Formal interpretation

for marginal parameters from unweighted GEE using odds ratio at union level can be done. For instance, for each unit increase in herd size and for artificially inseminated herds, on average, the odds of the prevalence of herd abortion is [*exp* (-0.4538)] = 0.635 times lower than that of the odds of herd used natural service breeding.

To identify the union-specific risk factors associated with brucellosis seropositivity and abortion, generalized linear mixed model (GLMM) were fitted to both responses individually. In the case of brucellosis seropositivity, only *herdsize* was significantly associated with the seroprevalence of brucellosis. The interpretation was that given the specific union, one unit increase in *herdsize*, the odds of being brucellosis seropositivity is [*exp (0.2245)*] = 1.25 times higher than the odds of being not seropositive in that union. Therefore, the seroprevalence of brucellosis in the herd increases with the increase in the size of the herd. The between union variance was estimated as 1.2127 (se = 0.4985) indicating that there were large differences between the unions. The intra-cluster correlation was 0.2693 (se = 0.0808) indicating that the herds within unions were significantly associated.

In the case of reported abortion, *breeding* and interaction effect with *herdsize* are highly associated with the herd prevalence of abortion. The negative interaction effect indicated that the artificially inseminated herd and larger herd size had lower probability of having prevalence of abortion. For instance, given the specific union, one unit increases in *herdsize* and artificially inseminated versus natural services breeding, the odds of the reported abortion is [*exp (-0.4820)*] = 0.618 times lower than the odds of not reported abortion in that union. In that case, the between union variance was estimated as 0.3806 (se=0.2732) indicating that there were also significant differences between the unions. The intra-cluster correlation coefficient indicated that the herds within the unions were also slightly correlated but not significant (*Table 10*).

So far, we have conducted different univariate approaches to find out the herd level risk factors associated with brucellosis seropositivity and abortion. Since we have two binary responses, it is of interest to know the correlation between reported abortion and brucellosis seropositivity in the herd and whether there is significant effect of the risk factors on either abortion or brucellosis seropositivity, the joint random effects model was fitted. The fixed effect parameter estimates

and the standard errors for both outcomes in the joint random intercepts model were same as the univariates random intercept model. The correlation parameter indicated that the association between *Brucella* seroprevalence and abortion were very low and was not statistically significant. Matope *et al*. (2011) showed in univariate case that there was close association between brucellosis seropositivity and abortion. Biologically, it was reported that infection of brucellosis causes abortion. Whereas Muma *et al*. (2007) reported at the herd-level that there was no association between brucellosis seropositivity and abortion. In the present study, we also did not find any relationship between abortion and brucellosis seropositivity. This finding was consistent with the culture and molecular diagnosis effort from aborted fetal membranes collected from the *Mymensingh* and *Sherpur* districts of Bangladesh.

4.2 Conclusion

Different approaches were fitted to identify the herd level risk factors associated with bovine brucellosis and abortion. Depending on the scientific question of interest, these approaches have different interpretations but in almost all the cases *breeding*, *herdsize* and their interaction effect were the most important herd-level risk factors associated with bovine brucellosis seropositivity and abortion in the present context of Bangladesh. In addition, there was no impact of brucellosis seropositivity on abortion.

4.3 Recommendation

- Since there is no brucellosis control program in Bangladesh, it should be implemented by the government and herd owners as well.
- Herds can be easily tested by bulk Milk Ring Test (MRT). All cattle of the positive herds should then be tested serologically by a highly sensitive and specific test like iELISA rather than SAT and RBT.
- Herd level vaccination may be started to reduce the prevalence level.
- Since no significant association between brucellosis and abortion was noted, further study should be explored to identify the causal agents responsible for abortion in Bangladesh.

References

1. Aerts, M., Faes, C., Hens, N., and Molenberghs, G. (2010) Handling Missingness and Informative Cluster Sizes When Modeling Prevalence and Force of Infection. *ENAR-JSM*, 3491-3501.
2. Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002) *Topics in Modeling of Clustered Data*. London: Chapman and Hall.
3. Ahasan, M.S., Rahman, M.S., and Hee-Jong, S. (2010) A Sero-surveillance of *Brucella* spp. Antibodies and individual risk factors of infection in cattle in Bangladesh. *Korean J Vet Serv,* **33(2)**, 121-128.
4. Amin, K.M.R., Rahman, M.B., Kabir, S.M.L., Sarkar, S.K., and Akand, M.S.I. (2004) Serological epidemiology of Brucellosis in Cattle of *Mymensingh* districts of Bangladesh. *Journal of Animal and Veterinary Advances,* **3(11)**, 773-775.
5. Aulakh, H.K., Patil, P.K., Sharma, S., Kumar, H., Mahajan, V., and Sandhu, K.S. (2008) A Study on the Epidemiology of Bovine Brucellosis in Punjab (India) Using Milk-ELISA. *ACTA VET. BRNO*, **77**, 393–399.
6. Bangladesh Bureau of Statistics (BBS). (2011) *Population and Housing Census 2011*. National Report (provisional), Dhaka, Bangladesh.
7. Banglapedia. (2011) *Mymensingh district*. National Encyclopedia of Bangladesh
8. Bennett, S., Woods, T., Liyanage, W.M., and Smith, D.L. (1991) A simplified general method for cluster-sampling surveys of health in developing countries. *World Health Stat. Q.,* **44 (3)**, 98–106.
9. Blood, D.C. and Radostits, O.M. (1989) *Veterinary Medicine: A Textbook of the Diseases of Cattle, Sheep, Pigs, Goats, Horses*. Balliere Tindall, London.
10. Cringoli, G., Rinaldi, L., Veneziano, V., Capelli, G., and Malone, J.B. (2002) A cross-sectional coprological survey of liver flukes in cattle and sheep from an area of the southern Italian Apennines. *Veterinary Parasitology,* **108**, 137–143.
11. Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G., and Bijnens, L. (2008) A High-Dimensional Joint Model for Longitudinal Outcomes of Different Nature. *Statistics in Medicine*, **27(22)**, 4408-4427.

12. Faes, C., Hens, N., Aerts, M., Shkedy, Z., Geys, H., Mintiens, K., Laevens, H., and Boelaert, F. (2006) Estimating herd-specific force of infection by using random-effects models for clustered binary data and monotone fractional polynomials. *Appl. Statist.* **55(5)**, 595–613.
13. Fieuws, S. and Verbeke, G. (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, **62**, 424-431.
14. Fieuws, S., Verbeke, G., and Molenberghs, G. (2007) Random effects models for multivariate repeated measures. *Statistical Methods in Medical Research*, **16(5)**, 387–397.
15. Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008) *Longitudinal Data Analysis*. Boca Raton: Chapman & Hall/CRC Press.
16. Fitzmaurice, G.M. (1995) A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data. *Biometrices*, **51(1)**, 309-317.
17. Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression.* New York: John Wiley and Sons, 2nd Edition.
18. Islam, A., Haque, M., Rahman, A., Rahman, M.M., Rahman, A., and Haque, F. (1983) Economic losses due to brucellosis among cattle in Bangladesh. *Bangladesh Veterinary Journal*, **17(1)**, 57-62.
19. Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
20. Matope, G., Bhebhe, E., Muma, J.B., Lund, A., and Skjerve, E. (2010) Herd-level factors for Brucella seropositivity in cattle reared in smallholder dairy farms of Zimbabwe. *Preventive Veterinary Medicine, Elsevier*, **94**, 213-221.
21. Matope, G., Bhebhe, E., Muma, J.B., Lund, A., and Skjerve, E. (2011) Risk factors for *Brucella* spp. Infection in smallholder household herds. *Epidemiol. Infect*, **139**, 157–164.
22. Matyas, Z. and Fujikura, T. (1984) Brucellosis as a world problem. *Dev. Biol. Stand,* **56**, 3-20.
23. Mia, A.S. and Islam, H. (1967) Preliminary study on the incidence of Bovine Infertility and the Economic loss caused by it. *Pakistan Journal of Veterinary Science,* **1**, 5-10.
24. Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.
25. Muma, J.B., Godfroid, J., Samui, K.L., and Skjerve, E. (2007) The role of Brucella infection in abortions among traditional cattle reared in proximity to wildlife on the Kafue flats of Zambia. *Rev. sci. tech. Off. int. Epiz.*, **26(3)**, 721-730.

26. Nahar, A. and Ahmed, M.U. (2009) Sero-prevalence study of brucellosis in cattle and contact human in *Mymensingh* district. *Bangl. J. Vet. Med*, **7(1)**, 269 – 274.
27. Nielsen, K. (2002) Diagnosis of brucellosis by serology. *Veterinary Microbiology*, **90**, 447-459.
28. OIE. (2000) *OIE Manual of Standards for Diagnostic Tests and Vaccines*. 4th edition, 12 Rue de Prony, 75017 Paris, France.
29. Omer, M.K., Skjerve, E., Woldehiwet, Z., and Holstad, G. (2000) Risk factors for Brucella spp. Infection in dairy cattle farms in Asmara, State of Eritrea. *Preventive Veterinary Medicine*, **46**, 257-265.
30. Pharo, H.J., Motalib, A., Alam, S., Fraser, G.C., and Routledge, S.F. (1981) Preliminary information on the prevalence of bovine brucellosis in the Pabna milk-shed area of Bangladesh. *Bangladesh Veterinary Journal,* **15**, 43-51.
31. Rahman, M.M. and Rahman, M.S. (1982) Study on the prevalence of brucellosis in cows in organized farms and domestic holdings in Bangladesh. *Bangladesh Veterinary Journal,* **16**, 53-58.
32. Roth, F., Zinsstag, J., Orkhon, D., Chimed-Ochir, G., Hutton, G., Cosivi, O., Carrin, G., and Otte, J. (2003) Human health benefits from livestock vaccination for brucellosis: case study. *Bull World Health Organ*. **81**, 867-876.
33. Speybroeck, N., Boelaert, F., Renard, D., Burzykowski, T., Mintiens, K., Molenberghs, G. and Berkvens, D. L. (2003) Design-based analysis of surveys: a Bovine Herpesvirus 1 case study. *Epidem. Infectn*, **13**, 991–1002.
34. Stringer, L.A., Guitian, F.J., Abernethy, D.A., Honhold, N.H., and Menzies, F.D. (2008) Risk associated with animals moved from herds infected with brucellosis in Northern Ireland. *Preventive Veterinary Medicine, Elsevier*, **84**, 72–84.
35. Vandeplassche, M. (1982) Reproductive efficiency in cattle: a guideline for projects in developing countries. *Animal Production and Health,* Paper No. 25. Food and Agriculture Organization of the United Nations (FAO), Rome.
36. Williamson, J.M., Datta, S., and Satten, G.A. (2003) Marginal analyses of clustered data when cluster size is informative. *Biometrics*, **59**, 36–42.
37. Young, J.E. (1995) An Overview of Human Brucellosis. *Clinical Infectious Diseases,* **21**, 283-290.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Herd-level Risk Factors Associated with Bovine Brucellosis Seropositivity and Abortion in Bangladesh

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Islam, Md. Atiqul

Datum: **16/09/2011**