Epidemiol. Infect. (2003), **131**, 991–1002. © 2003 Cambridge University Press
DOI: 10.1017/S095026880300102X Printed in the United Kingdom

Design-based analysis of surveys: a bovine herpesvirus 1 case study

N. SPEYBROECK[1]*, F. BOELAERT[2], D. RENARD[3], T. BURZYKOWSKI[3], K. MINTIENS[2], G. MOLENBERGHS[3] AND D. L. BERKVENS[1]

[1] *Institute for Tropical Medicine, Nationalestraat 155, 2000 Antwerp, Belgium*
[2] *Co-ordination Centre for Veterinary Diagnostics, Veterinary and Agrochemical Research Centre, Groeselenberg 99, 1180 Brussels, Belgium*
[3] *Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, 3590 Diepenbeek, Belgium*

(*Accepted 6 October 2002*)

SUMMARY

This paper critically assesses the design implications for the analysis of surveys of infections. It indicates the danger of not accounting for the study design in the statistical investigation of risk factors. A stratified design often implies an increased precision while clustering of infection results in a decreased precision. Through pseudo-likelihood estimation and linearisation of the variance estimator, the design effects can be taken into account in the analysis. The intra-cluster-correlation can be investigated through a logistic random effect model and a generalised estimating equation (GEE), allowing the investigation of the extent of spread of infections in a herd (cluster). The advantage of using adaptive Gaussian quadrature in a logistic random effect model is discussed. Applicable software is briefly reviewed. The methods are illustrated with data from a bovine herpesvirus 1 (BHV-1) serosurvey of Belgian cattle.

INTRODUCTION

An important aspect of veterinary epidemiology is the quantitative investigation of disease occurrence [1]. This includes surveys, monitoring, and surveillance that often deal with binary data. Three important characteristics of survey data are stratification, clustering and sampling weights.

Stratification is a sampling method aimed at reducing variance, when a known factor causes significant variation in the outcome variable but is not the target of analysis. For example, in the case of beef production in a population of two different breeds, sampling variation of estimates will be substantial, largely due to genetic differences affecting beef production between the two breeds. Stratification by breed will allow reduction of the overall variation in the beef production estimate. The technique also allows easy access to information about the sub-populations represented by the strata. For stratified sampling to be effective at reducing variation, the elements within the strata should be homogenous and variance between the strata should be large. A possible disadvantage, however, is that the status of the sampling units with respect to the stratification factor must be known and more complex methods are required to obtain correct variance estimates [2].

Clustering of data may be due to repeated measurements of subjects over time or due to sub-sampling of the primary sampling units. Livestock disease clustering is a consequence of unequal distribution of the disease agents throughout the animal population [3]. Interest primarily concerns individual-level

* Author for correspondence.

characteristics, such as the disease status of the animal, but the sampling unit becomes a grouping of individual animals such as the herd to which they belong. The groups or clusters can represent natural groupings such as litters or herds, or they can be based on artificial groupings such as geographic areas or administrative units. The random selection of the clusters as the sampling units can be performed using simple random, systematic or stratified random sampling. Clustering can help to reduce the sampling and data collection costs. However, since independence among sample observations is a key assumption underlying standard statistical procedures, the presence of clustering in the data may raise important statistical issues, which should be addressed in the analysis. With data collected on the basis of clusters, the variance is largely influenced by the number of clusters, not the number of animals in the sample. Cluster sampling can lead to an increased sampling variance, in which case a larger sample size would be required to reduce the variance to acceptable levels [2]. Unfortunately, scientific reports often present, for example, confidence intervals that assume simple random sampling whereas the design involved clustered units.

In sample surveys observations are selected through a random process, but different observations may have differing selection probabilities. These probabilities should be accounted for in the analysis of the survey, otherwise biased estimates may be obtained [4].

The objective of this paper is to present methods for analysing survey data where clustering, stratification and differing sampling probabilities may be present. The emphasis is directed towards the estimation of the effect of risk factors on the presence of infectious diseases. As the main tool, logistic regression models, taking into account the effects of clustering and stratification, are considered. Furthermore, ways of calculating the intra-cluster correlation are presented. Various software packages, which can be used to apply these methods of analysis, are considered. The effect of ignoring sampling design characteristics is demonstrated and discussed. The techniques are illustrated using data from a bovine herpesvirus 1 (BHV-1) serosurvey of Belgian cattle, reported by Boelaert et al. [5]. BHV-1 causes infectious bovine rhinotracheitis, an enzootic disease. Programs to eradicate BHV-1 have been implemented in several European countries to facilitate the free trade of cattle within the European Union.

METHODS

The data

Figures 1a–d present the design of the BHV-1 serosurvey of the Belgian cattle population. The survey was conducted on cattle herds of all types from December 1997 to March 1998 in Belgium. The sample was stratified by province (ten provinces in Belgium). Within each province, 1% of the total number of herds was sampled. In the selected herds, all animals were blood sampled. The sera were tested for antibodies against BHV-1 by using a commercially available blocking ELISA (HerdCheck®, Idexx, France), specific for BHV-1 glycoprotein B [6]. The age and sex of 11 284 animals originating from 309 BHV-1 unvaccinated herds were collected. Also, the type (dairy, mixed or beef) and size of the herds were registered. Due to computational problems, the variable 'herd size' was dichotomised as 0 (or 1) for farms smaller (or larger) than the average herd size ($=36{\cdot}5$) in the final models presented in Section 3.

The survey is an example of a one-stage cluster sampling design. The individual subjects (animals) still remain the target units so that animal-level disease can be studied, but the primary sampling unit becomes a group of individuals (the herd). All elements within a randomly selected group are included in the sample. This technique requires a sampling frame for the groups, but not for the members within the groups. In the present example, the random selection of clusters (herds) as the sampling units was performed using stratified random sampling, but it can also be performed using systematic or simple random sampling [2].

Important features of survey data

In this section, the following features of survey data are shortly discussed: stratification, clustering and sampling probabilities.

Stratification

The argument in favour of stratification can be illustrated using a simple example. Consider a population of three male animals, all of which are infected, and three females that are not, then clearly the prevalence of the disease is 50%. Now, assume that from the above population independent samples of size 2 are to be taken using simple random sampling without replacement. Within the above population, the

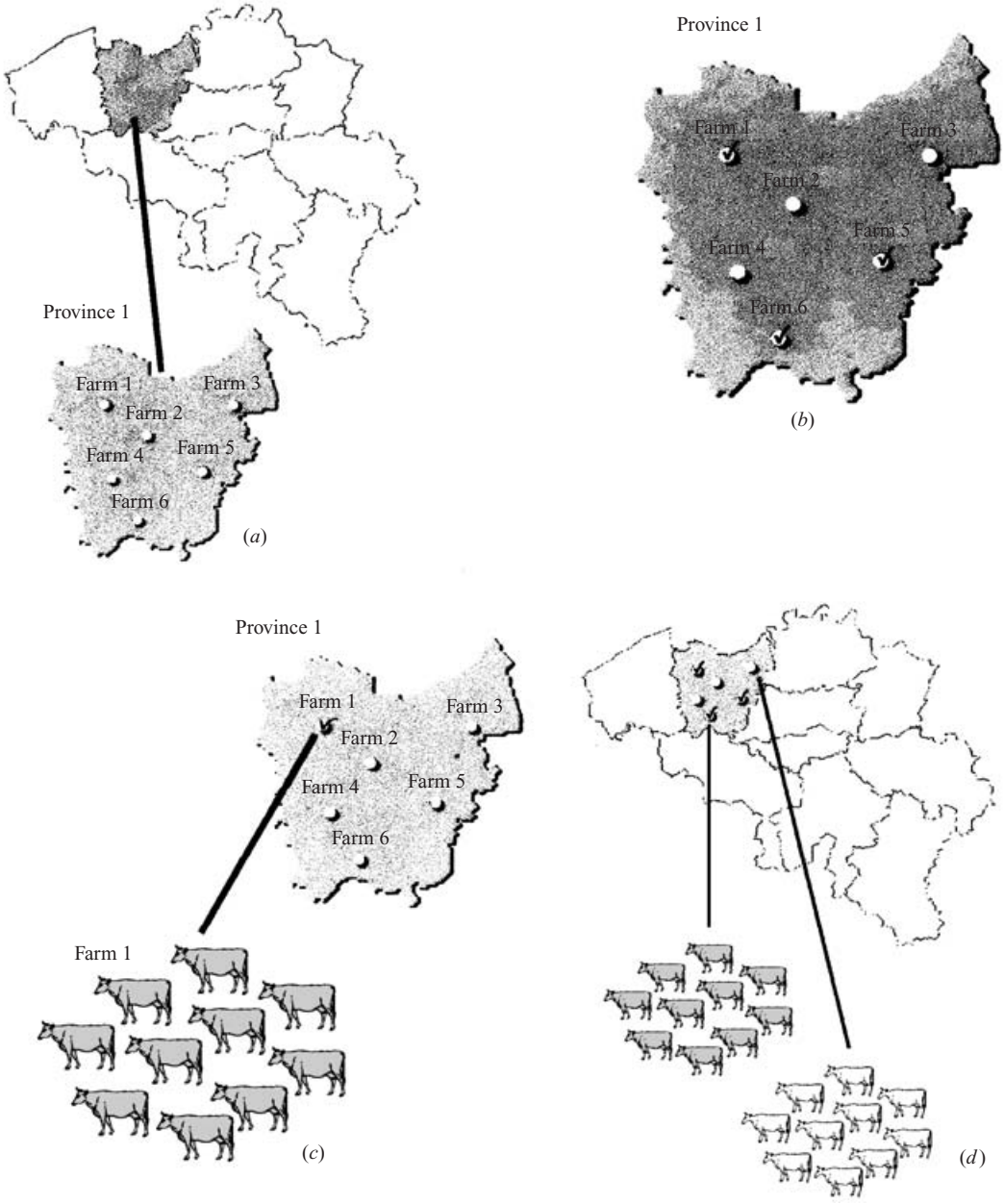

Fig. 1(*a–d*)**.** Design of the BHV-1 serosurvey of the Belgian cattle population, 1998. (*a*) In every province the farms are listed. (*b*) A same proportion of the farms is selected in each province. (*c*) In each selected farm all the animals are sampled. (*d*) A total overview of the survey: a same proportion of the farms is selected in each province and in each selected farm all the animals are sampled.

probability of estimating prevalences as 0% is 0·2, as 50% is 0·6 and as 100% is 0·2. As a consequence there is a probability of 0·4 that the estimated prevalence differs markedly from the true value. However, if we consider sampling stratified by sex, with one animal sampled from males and one from females, the prevalence will be estimated as 50% for every sample, the variability of the estimate thus being greatly reduced.

In real settings, stratification is an effective method for reducing variability of an estimator if a known factor, which is not the target of analysis, causes substantial variation in the outcome variable such that the elements within the strata are homogeneous and variability between the strata is large. Stratified sampling also leads to a straightforward computation of estimates for the sub-populations represented by the strata. Obviously, the technique requires that the

status of the sampling units with respect to stratification factors be known. Also, more complex methods are required to obtain precision estimates referring to the global population [2].

Clustering

Clustering can be seen as a form of stratified design, where instead of selecting some individuals from each (large) stratum, we aim at selecting some (from a large number of relatively small) strata with, possibly, each individual within a selected group included in the sample [7]. To reflect this difference, the strata are called clusters. If all individuals in a sampled cluster are included in the sample, as in the BHV-1 survey, this is referred to as one-stage cluster sampling.

If a disease is contagious, the clustering (or grouping) of animals within herds may result in a higher chance for an animal becoming infected once the infection is introduced into the herd. Consequently, individual responses, i.e., whether the animals are infected or not, are more homogeneously distributed within herds than in the whole population. If the response is distributed in a homogeneous way within a cluster, considering the sample as a simple random sample can lead to erroneous conclusions. Assume for example the extreme case of herds in which either all or none of the animals are infected. In this situation, the calculation of the standard errors of the parameter estimates should be based on the number of farms rather than on the number of animals since the information provided by a single animal would amount to the total information provided by the whole herd to which the animal belongs.

Consequently, in the presence of clustering, the calculation of the variance of the prevalence using formulae for simple random sampling may yield overly optimistic estimates. Unfortunately, it appears that in surveillance studies investigating prevalence of diseases the precision of the prevalence is often overestimated [8].

Sampling probabilities

As a result of the choice of a sampling design, each individual member of a population is assigned a probability with which it can be included in the sample. If these probabilities differ between members they should be accounted for in the analysis of the survey. Consider the BHV-1 survey. Since in the sampled herds, all animals were included in the sample, the probability of being sampled was the same for each animal in the population, as the herds were sampled with equal probabilities. Suppose, however, that from each of the sampled herds only one animal would have been sampled. (In such a case the sampling would have been an example of two-stage sampling, with herds considered as *primary sampling units* and the animals as *secondary sampling units*.) Then the probability of being selected would have been higher for animals from small herds than for those from large herds. As a result, the animals from small herds would have been over-represented in the sample. It follows that the estimates obtained from the analysis ignoring the sampling probabilities would have been biased towards the characteristics of the sub-population of small-herd animals. This was not the case for the BHV-1 survey.

The analysis of a survey can be adjusted for unequal sampling probabilities by applying appropriate *weights* to the observed results [4]. In general, different sampling probabilities arise most naturally in *multi-stage sampling designs*. Such designs, in combination with stratification, are recommended by the Office International des Epizooties as part of the official pathway to declaration of freedom from infection with the rinderpest virus.

In short, adjusting for unequal sampling probabilities allows unbiased point estimates to be obtained from survey data. Taking into account clustering and stratification results in appropriate precision measures for the point estimates. Adjusting for sampling probabilities can also influence the precision.

Statistical methodology

We will consider the situation of a binary response variable Y. In our example, Y will indicate whether an animal is infected ($Y=1$) or not ($Y=0$). To investigate the effect of explanatory variables (age or sex of an animal, for example) on the probability of infection, we will consider a logistic regression model. Denoting as π(covariates) the probability of infection as a function of covariates, we can write the model symbolically as follows:

$$\begin{aligned}\text{logit}\{\pi(\text{covariates})\} &= \log\{\pi(\text{covariates})/ \\ &[1-\pi(\text{covariates})]\} = \alpha + \beta^{*}\,\text{covariates} \qquad (1)\end{aligned}$$

The parameters α and β (which is a vector) have to be defined.

For the BHV-1 survey, age and sex of the animals, as well as type (dairy, mixed, beef) and size of the herd, will be used as covariates.

Adjustment for sampling probabilities and stratification

In general, this is done by appropriate *weighting* of the data. The weights are taken as the inverse of the sampling probabilities [9, 10]. Here, we will not discuss the techniques any further; the interested reader can find them in any textbook on survey sampling [11–13].

Using a 'pseudo' likelihood [9] and deriving the variance estimator through 'linearisation' is one way to account for the effects of sampling probabilities and stratification as well as clustering (see next section) in the analysis. A pseudo-likelihood is needed since the standard ML estimator does not give a true 'likelihood' under a complex design and therefore estimates of model parameters are obtained by solving weighted analogues of likelihood equations based on the probability sampled data. The full MLE would require an expression for the exact likelihood, which may be very complicated and require many assumptions since it involves modelling the relationship between the response and the design variables [9]. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables one to use statistical theory to make valid inferences from the sample to the survey population. As a consequence of these issues, the likelihood-ratio test is invalid with weighted data of this kind.

Adjustment for clustering

There are many methods available for the analysis of clustered binary data. In general, one can distinguish between marginal, conditional and random-effects approaches, which can be applied using different inferential methods (likelihood, quasi- or pseudo-likelihood, generalised estimating equations). Unlike the Gaussian setting, they tend to give dissimilar results. Reviews can be found in Diggle, Liang and Zeger [14], Fahrmeir and Tutz [15] or Pendergast [16]. We will discuss only the approaches used most frequently in the survey context.

The analysis at herd level

In an attempt to account for clustering of the animals, one may consider analysing the data at herd level. In such an analysis, a herd with at least one seropositive animal is called positive, otherwise it is called negative. A logistic regression with binary response Z, indicating whether a herd is positive ($Z=1$) or not ($Z=0$), can be carried out. Only herd-level covariates can be used in such a model, for example, the type and size of the herd, the average age of the animals and the proportion of males by herd. A herd-prevalence could be defined as the proportion of herds with at least one positive animal.

An advantage of the herd-level analysis is that it focuses on the probability of infection in a herd, which is economically important information. However, there are several disadvantages. As already mentioned, only herd-level covariates can be considered in the analysis. Moreover, the associations detected at the herd level do not necessarily correspond to those existing at the animal level. Thus, there might be some confusion between aggregate and individual effects, an issue that is often referred to as the ecological fallacy [17].

Marginal model fitted using generalised estimating equations

One way to address the disadvantages associated with the herd-level analysis is to fit the logistic regression model (1), while correcting estimated standard errors of parameters β for clustering. The approach can be applied using the generalised estimating equations (GEE) technique developed by Zeger and Liang [18].

Logistic random-effects model

The use of a random-effects model approach (see Agresti et al. [19] for a recent review in the broader context of categorical response data) can be motivated by arguing that animals belonging to a herd share the same environment (physical location), as well as characteristics such as the type of farm (milk- or meat-oriented). These shared factors, whose effects can change from herd to herd, create dependencies between responses observed for the individual animals.

In its simplest form, the model can be symbolically written as follows (with b_i random variable representing the effect of factors shared by the animals belonging to herd i):

$$\text{logit}\{\pi(\text{covariates}, b_i)\} = \alpha + \beta^* \text{covariates} + b_i, \qquad (2)$$

where i is an index for herds, $\pi(\text{covariates}, b_i)$ denotes the conditional probability of infection (conditionally on the covariates and the random effect b_i). Usually, these random variables are assumed to be normally distributed. Likelihood inference in this type of model

can proceed by integrating over the random effects b_i to derive the marginal likelihood, which can practically be done by numerical integration (Gaussian quadrature for instance). Note that the interpretation of the β coefficients in this model is conditional on the (unobserved) value of the random variable b_i and is, therefore, called 'individual-specific'. In model (1), on the other hand, β can be interpreted as describing marginal (so-called 'population-averaged') effects of the covariates. This model does not take into account clustering and other design effects.

It should also be noted that model (2) assumes that, conditionally on the value of b_i, the response has a binomial error distribution. A different kind of 'random-effects' model would be to assume that conditional on each herd, the response is binomial and that the response probabilities follow, for instance, a β distribution (thus yielding the so-called β-binomial model). We will not illustrate this approach in the present paper. A description of the approach can be found in Kleinman [20].

Comparison with pseudo-likelihood methodology

Unlike pseudo-likelihood methodology [4], standard marginal and random-effects models cannot directly account for survey-related issues but simply afford a more flexible way to account for clustering in the analysis of the data. Note, however, that for the data at hand, one way to take the stratification into consideration would be to incorporate strata indicators as covariates.

Quantifying the influence of sampling design on the precision: design effect

The influence of sampling design on the precision of estimates can be quantified using the measure of design effect [13]. It is defined as

$$deff = \hat{V}/\hat{V}_{\text{srswor}} \tag{3}$$

where $\hat{V}$ is the design-based estimate of the parameter variance and $\hat{V}_{\text{srswor}}$ is an estimate of the variance for a hypothetical simple-random-sampling design in place of the complex design that was actually used. $\hat{V}$ can be computed by adapting the formulae for variance estimation by using techniques such as (balanced) replicated sampling, jackknife repeated replication and the Taylor series method [21].

In cases where the quantity of interest is a proportion (such as the prevalence discussed so far) the design effect is directly proportional to the size of the intra-cluster-correlation and the cluster size [22]:

$$deff = 1 + \rho(m-1), \tag{4}$$

where m denotes the cluster size (assumed to be constant) and ρ is the intra-cluster correlation. For a variable cluster size, a reasonable approximation for m is the average cluster size.

Marginal logistic regression models fitted using GEE and random-effect models allow ρ to be estimated. Apart from indicating the amount of association between responses within a cluster, the intra-cluster correlation can be interpreted as measuring the part of the total variance explained by the clusters [13, 23].

In the following section, software, which can be used to carry out the aforementioned models, is presented.

Software

There are several commercially available statistical software packages where methods for analysing survey data are implemented: STATA, SAS and SUDAAN, amongst others. Since the last was not available to us, we will only consider STATA and SAS.

STATA

STATA (Stata Corporation, Texas, USA) is a multipurpose interactive statistical package. It includes a set of so-called *svy* commands implementing methods for analysing surveys.

A pseudo-likelihood method is used to calculate estimates of model parameters by solving weighted analogues of likelihood equations based on the probability sampled data. For all models, STATA can adapt standard errors and confidence intervals for estimates taking into account clustering, stratification and sampling probabilities. All *svy* commands can compute design effects for their estimates. STATA currently uses the Taylor series linearisation estimator as the variance estimation method. This method approximates the complex formula for the adapted variance, by writing it as a series of functions. Lee et al. [21] give a detailed account of this method. Multistage designs are handled by the 'ultimate' cluster sample selection paradigm [4]. Kish [13] defines the term ultimate clusters. The ultimate cluster is a grouping of sampled cases for variance estimation purposes. In general, the use of ultimate clusters for sampling error estimation reflects the gains in precision from stratification and the loss in precision

from the clustering of cases within primary sampling units. Under the ultimate cluster-sampling model, elements within primary sampling units are divided into ultimate clusters and a sample of these clusters is drawn without replacement across the primary sampling units. Variance estimates are computed using only between first stage unit totals without having to compute the variance components at each stage of selection.

Svylogit fits logistic regression models for survey data and is able to incorporate probability sampling weights, stratification and clustering (one level only) or any combination of these three. Associated variance estimates and design effects (*deff*) are computed. Clustering is taken into account by pseudo-likelihood estimation and precision estimates are calculated in a way similar to the sandwich estimator in GEE.

The *xtlogit* command can fit marginal (using GEE) as well as random-effects logistic regression models. In the latter case, numerical integration is based on simple Gaussian quadrature, although a limitation of the command is that it cannot handle more than 30 quadrature points.

SAS

SAS (SAS Institute Inc., North Carolina, USA) is also a multi-purpose statistical package. From version 7 of the SAS system onwards, some procedures have been available for the analysis of data from complex sample surveys. In particular, the SURVEYSELECT procedure selects probability samples using various sampling designs; the SURVEYMEANS procedure computes descriptive statistics for sample survey data; and the SURVEYREG procedure fits linear regression models for such data.

Two other procedures are of interest in the present context. These are the GENMOD procedure, which can fit models for correlated data using the GEE method, and the NLMIXED procedure, which fits nonlinear mixed models, that is, models in which both fixed and random effects enter nonlinearly. NLMIXED can, in particular, handle logistic random-effects regression models and use adaptive Gaussian quadrature to approximate the likelihood, which are preferable to simple Gaussian quadrature in general [24].

RESULTS

In this section, we illustrate the issues previously discussed using the data from the BHV-1 survey. At first, the data are analysed at herd level and this is compared with a naïve analysis at the animal level. Thereafter, the data are analysed at the animal level, using the more sophisticated methods described in Section 2.

Table 1. *Average seroprevalence of animals and herds, by type of farm, BHV-1 serosurvey, Belgium, 1998*

	Seroprevalence	
Type of farm	Animal level ($n=11\,284$)	Herd level ($n=309$)
Dairy	0·345	0·864
Mixed	0·429	0·907
Beef	0·320	0·541

Table 2. *Frequency distribution of farm size per type of farm, BHV-1 serosurvey, Belgium, 1998*

Size (number of animals)	Dairy farms	Mixed farms	Beef farms
$\leqslant 20$	6	5	152
$\leqslant 40$	8	5	21
$\leqslant 60$	16	19	8
$\leqslant 80$	5	8	7
$\leqslant 100$	9	8	2
$\leqslant 120$	8	4	1
$\leqslant 140$	4	2	0
$\leqslant 160$	1	0	0
$\leqslant 180$	1	1	1
>180	1	2	2
Total	59	54	194

Analysis at the animal level *vs.* analysis at the herd level

The herd seroprevalence was estimated at 67% (207 of 309 farms had at least one animal infected). The estimated proportion of seropositive animals was 36%, which confirms the figure of Boelaert et al. [5]. Table 1 shows the average seroprevalence for animals and herds by type of farm. Interestingly, beef farms show a much lower herd seroprevalence. The lower seroprevalence in beef farms can be explained by the fact that these farms are typically small as shown in Table 2, but carry the same weight in the calculation of the herd-seroprevalence. As the prevalence tends to increase with size, the large number of small beef farms results in a lower herd-prevalence for this type of farm. This illustrates how size of the farm acts as

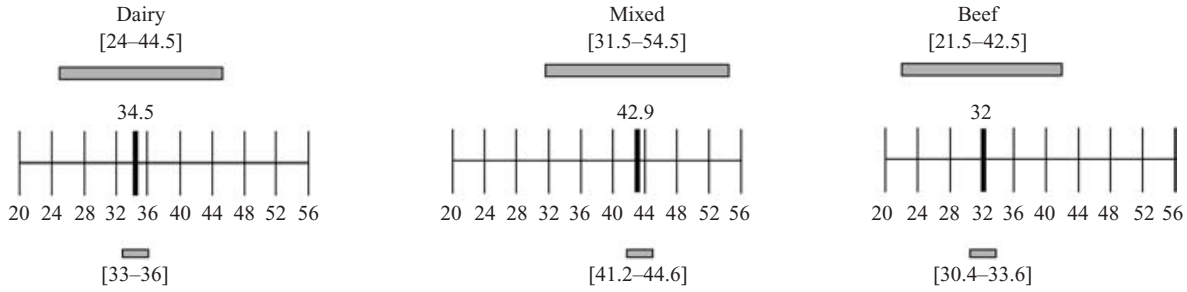

Fig. 2. BHV-1 seroprevalence by type of farm: estimates and 95% confidence intervals, respectively taking into account (confidence intervals as top-bar) and not taking into account (confidence intervals as bottom-bar) the effects of the sampling design.

a confounding factor in the analysis at herd level. This was confirmed by the logistic regression at the herd level, which indicated that beef farms had a significantly ($P<0{\cdot}001$) lower prevalence than the other types of farms. However, including size of the farm in the model showed that a larger size was significantly ($P<0{\cdot}001$) related to a higher risk of herd seroprevalence and that beef farms had no longer a significantly lower prevalence than the other types of farms ($P=0{\cdot}331$).

At the animal level, the difference in seroprevalence between the different types of farms was not so pronounced. This can be explained by the stronger weight of big farms in the final seroprevalence calculations. The seroprevalence at animal level in large farms is close to the figures obtained for animal level in Table 1.

Effects of the sampling design on confidence intervals

Figure 2 presents, by type of farm, the estimated seroprevalence at animal level together with 95% confidence intervals (in brackets) computed with and without the effects of clustering and stratification having been taken into account. The results were obtained using the command *svymean* in STATA. Clearly, if the sampling design is not accounted for, the precision of the estimates is considerably overestimated.

For the data at hand, it is interesting to investigate which of the sampling design features, (stratification or clustering) is mainly responsible for the adjustment of the confidence intervals. The estimates of BHV-1 seroprevalence by province (proportion of animals infected by province) ranged from 2 to 42·2%. A strong difference between provinces indicate that including this factor in the analysis could improve the estimation in terms of efficiency. However, with the command *svymean* of STATA the design effect for stratification was only 0·96, while the design effect for clustering was 45·6, which is very similar to the design effect for clustering and stratification together (45·5).

A similar effect of adjustment for sampling design on the precision estimates can be seen when considering a logistic regression model with type and size (binary variable) of the farm, and age and sex of the animal, included as co-variates. Table 3 presents the coefficients of the model as well as estimates of their standard errors obtained using the STATA command *svylogit* with and without adjustment for clustering and stratification. Without adjusting for sampling design, Table 3 (misleadingly) indicates that the seroprevalence among animals in mixed farms is significantly higher than in dairy and beef farms. Also, the naïve model suggests significant effects of size of the farm, age and sex of animals.

When sampling design is accounted for, the results change substantially, owing to the increase in standard errors of the coefficients. Conclusions for the herd-level covariates are mostly affected: the effect of mixed farms becomes non-significant, while the effect of size of the farm becomes borderline significant at the 5% level.

As in the case of the overall BHV-1 seroprevalence, design effect associated with adjusting the logistic regression model for stratification was small (0·96). It seems that the homogeneity in the strata was not high enough compared to the variability between the strata.

Marginal GEE model

Using the same covariates as in the previous model, a marginal model (GEE with exchangeable working correlation structure) was fitted with the *xtlogit*

Table 3. *Effect of type and size of the farm, and age and sex of the animal, on the BHV-1 seroprevalence in Belgian cattle, with and without adjustment for clustering and stratification*

		Without design effect		With design effect	
Variable	Coefficient	S.E.	$P>\|t\|$	S.E.	$P>\|t\|$
Herd size (binary)	0·691	0·066	0·000	0·283	0·015
Age	0·215	0·011	0·000	0·024	0·000
Sex	0·324	0·067	0·000	0·153	0·035
Mixed	0·348	0·048	0·000	0·332	0·296
Beef	0·072	0·056	0·201	0·365	0·845

Sex: 0 = female, 1 = male; Herd size (binary) = 0 if herd size $\leqslant$ 36·5, and = 1 otherwise.

Table 4. *Estimates with standard errors and significance level for a marginal model, BHV-1 serosurvey, Belgium, 1998*

Variable	Coefficient	S.E.	$P>\|t\|$
Herd size (binary)	0·838	0·245	0·001
Age	0·214	0·007	0·000
Sex	0·137	0·050	0·006
Mixed	−0·365	0·276	0·187
Beef	−0·285	0·275	0·300

Table 5. *Estimates with standard errors and significance level for a random-effects logistic model, BHV-1 serosurvey*

Variable	Coefficient	S.E.	$P>\|t\|$
Herd size (binary)	1·024	0·416	0·014
Age	0·424	0·016	0·000
Sex	0·301	0·100	0·003
Mixed	0·284	0·478	0·553
Beef	−0·069	0·472	0·884

command in STATA and the results are shown in Table 4. The value of the intra-herd correlation coefficient was estimated as 0·526.

To further account for omitted aspects of the sampling design, we could, in principle, incorporate indicator variables for provinces in the above model but this led to convergence problems.

Random-effects logistic regression

With simple Gaussian quadrature, the estimated coefficients and their standard errors were strongly dependent on the number of quadrature points specified in the algorithm. This is a typical sign that Gaussian quadrature works poorly and that adaptive Gaussian quadrature should be utilised instead. In fact, the latter should be preferred as it has, in general, better numerical properties than simple Gaussian quadrature. For this reason, we illustrate the fitting of the random-effects model solely with the SAS procedure NLMIXED. The results are shown in Table 5.

The value of σ^2, the between-herd variance, was estimated as 5·964 (S.E. = 0·732). From this model, the derived value of the intra-herd correlation coefficient can be calculated as:

$$\rho = \frac{\sigma^2}{\pi^2/3 + \sigma^2},$$

which gives 0·645 (S.E. = 0·028).

Model comparison

The conclusions reached by the pseudo likelihood model, the marginal model and the random effect model are similar: Herd size, age and sex are significantly associated with the seroprevalence of BHV-1 at the 5% level. The signs of the coefficients differ only for non-significant parameters: the type of the farm. The intra-cluster correlation is above 50% with a marginal model as well as with a random effect model.

DISCUSSION

In general, while adjusting for the sampling design effects in a complex survey, two approaches might be considered to analyse data from a survey with (clustered) binary responses indicating, for example whether animals are seropositive or not. First, one

might consider clusters (herds) as the units of analysis. Information at the herd-level may be sufficient when one aims to eradicate disease from the herd. This is particularly true for BHV-1, because control and eradication measures implicate the herd, not the animal [25]. In the analysis at herd level of the BHV-1 survey, the smaller size of the majority of the beef farms resulted in a lower herd seroprevalence for beef farms. This was due to the confounding effect of the size of the farm, and demonstrates the importance of correcting for such covariates. Secondly, one might treat animals as the units of analysis and adjust the results for the effects of clustering, stratification and weighting. Such an analysis at animal level offers an additional and vital insight into the epidemiology, since analysis only at herd level does not allow consideration of the factors measured at animal level. The adjustment for the effects of clustering, stratification and weighting can be done using pseudo-likelihood methods with a linearisation estimator for the variance.

Although a BHV-1 risk factor study was outside of the scope of this paper, the influence of four different parameters on the serological results was investigated with the sole aim of illustrating the impact of the design on the analysis; type and size (binary variable) of the farm, and age and sex of the animal. Size of the farm and age of the animals seemed to be statistically important predictors for BHV-1 seropositivity of an animal. The older an animal is, or the larger the farm it belongs to, the higher is the chance of it having a positive serological result. Both risk factors can be explained from a biological viewpoint. The risk of BHV-1 transmission among cattle within herds is higher in larger herds. This may be explained by the within-herd contact structure. In smaller herds the number of susceptible animals is smaller throughout the year, so infection may not be sustained. Larger herds usually have loose-housing barns, creating more contact between infected and susceptible animals. These herds possibly also have more visits by animal handlers (farmers, inseminators, veterinarians, traders) [26]. Also the life cycle of herpesviruses provides a further explanation of herd size being a risk factor. During primary infection, herpesviruses are disseminated within susceptible populations, which raise strong immune responses and overcome in most cases the diseases associated with the infections. The latent viruses represent a long-term reservoir that becomes epidemiologically meaningful upon reactivation. Then, seemingly healthy animals are able to re-excrete and transmit the virus to non-immune as well as to immune hosts [27–29].

The advantage of the pseudo-likelihood methods is that they take account of all the design effects simultaneously. However, they do not provide an estimated intra-cluster correlation and do not allow for more than one level of clustering in STATA. Variance estimation for multistage sample data is carried out through the customary between-primary sampling units-differences calculation. The generalised estimating equations and the random-effects logistic regression model that can be carried out with respectively the commands GENMOD and NLMIXED in SAS cannot directly take into account stratification. However, these commands are not specially designed to analyse survey data and by using the stratum as a fixed effect in the model, the stratification variable is accounted for in a somewhat less efficient way.

The sampling design must be taken into account when analysing surveys. Otherwise, wrong conclusions may be drawn. In general, stratification may lead to an increase of the precision of the estimates, while clustering may decrease the precision. This was illustrated using the pseudo-likelihood methods with a Taylor series linearisation estimator as the variance estimation method. The type of farm can erroneously be considered as significantly related to the seroprevalence by not including an adjustment for the design. Adjusting for unequal sampling probabilities results in unbiased point estimates. Accounting for stratification and clustering allows correct standard errors to be made. Thrusfield [1] stated that the group of animals which is of most epidemiological importance in terms of the transmission and maintenance of infection, and therefore for disease control and eradication, is the herd. In the above methods for analysis at animal level, the herd was still considered as the primary sampling unit.

It should be mentioned that SUDAAN is also suitable for the analysis of data from complex sample surveys. In particular, in addition to the Taylor series linearisation as a robust variance estimation method, it allows also for the use of replication methods.

The precision of the estimates for the data of BHV-1 was slightly increased by including the effect of stratification at the province level. However, this had only a minor effect (0·96, with 1 meaning no effect) on the parameter estimates, which might be explained by the variability within provinces not being smaller than the overall variability. The effect of

clustering was much stronger, as could be seen from the widened confidence intervals (pseudo-likelihood approach) or by the relatively high values obtained for the intra-herd correlation coefficient in the marginal and random-effects model. Also, the sampling design effects influenced mostly the herd-level (e.g. type of farm) rather than the animal-level covariates (e.g. age of the animal). Cattle in Belgian farms are kept together in lots. This constitutes conditions for the infection to spread, which results in more homogeneous clusters with respect to the presence or absence of the infection. This was supported, in the analysis, by the value of the intra-cluster correlation, which was higher than 50%. As an alternative to using pseudo-likelihood methods, logistic random effect models were used. The importance of using adaptive Gaussian quadrature for the random effect models was illustrated. By not using it, different conclusions with respect to the significance and signs of the coefficients are reached depending on the number of quadratures used.

Additionally, the testing procedure had inherent probabilities of misclassification, due to diagnostic test inaccuracy. If the diagnostic sensitivity and specificity of a test are known, the true prevalence can be estimated. Unfortunately, the test characteristics are, in general, not known. Moreover, these test characteristics vary among sub-populations [30]. The impact of this misclassification on the BHV-1 risk factor analysis is currently being investigated. In conclusion, cross-sectional surveys based on diagnostic test results will always be concealed by the inaccuracy of the diagnostic test and this cannot be solved by a complex analysis.

ACKNOWLEDGEMENTS

We thank Dr Maxime Madder, Institute for Tropical Medicine, Antwerp, Belgium for skilled technical assistance. The BHV-1 survey was supported by the Fund for Animal Health and Production, Ministry of Small Enterprises, Traders and Agriculture, Belgium.

REFERENCES

1. Thrusfield M. Veterinary epidemiology. 2nd ed. Cambridge, United Kingdom: Blackwell Science Ltd, 1995.
2. Pfeiffer DU. Veterinary epidemiology – an introduction. Royal Veterinary College, University of London, United Kingdom, 1999.
3. Rothman KJ. A sobering start for the cluster busters' conference. Am J Epidemiol 1990; **132**: S6–S13.
4. StataCorp. Stata Statistical Software: Release 7.0. College Station, TX: Stata Corporation, 2001.
5. Boelaert F, Biront P, Soumare B, et al. Prevalence of bovine herpesvirus 1 in the Belgian cattle population. Prev Vet Med 2000; **45**: 285–295.
6. Kramps JA, Magdalena J, Quak J, et al. A simple, specific, and highly sensitive blocking enzyme-linked immunosorbent assay for detection of antibodies to bovine herpesvirus 1. J Clin Microbiol 1994; **32**: 2175–2181.
7. Barnett V. Sample survey principles and methods. London: Edward Arnold, 1991.
8. McDermott JJ, Schukken YH, Shoukri MM. Study design and analytic methods for data collected from clusters of animals. Prev Vet Med 1994; **18**: 175–191.
9. Skinner CJ, Holt D, Smith TMF, eds. Analysis of complex surveys. Chichester: Wiley, 1989.
10. Sarndal C-E, Swensson B, Wretman J. Model assisted survey sampling. New York: Springer, 1992.
11. Cochran WG. Sampling techniques. 3rd ed. New York: Wiley, 1977.
12. Levy PS, Lemeshow S. Sampling of populations. 3rd ed. New York: Wiley, 1999.
13. Kish L. Survey sampling. New York: Wiley, 1965.
14. Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. Oxford: Clarendon Press, 1994.
15. Fahrmeir L, Tutz G. Multivariate modelling based on generalized linear models. New York: Springer-Verlag, 1994.
16. Pendergast JF, Gange SJ, Newton MA, Lindstrom MJ, Palta M, Fisher MR. A survey of methods for analyzing clustered binary response data. International Statistical Review 1996; **64**: 89–118.
17. Robinson WS. Ecological correlations and the behavior of individuals. American Sociological Review 1950; **15**: 351–357.
18. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; **42**: 121–130.
19. Agresti A, Booth JG, Hobert JP, Caffo B. Random effects modeling of categorical response data. Sociological Methodology 2000; **30**: 27–80.
20. Kleinman JC. Proportions with extraneous variance: single and independent samples. J Am Statist Assoc 1973; **68**: 46–54.
21. Lee ES, Forthofer RN, Lorimor RJ. Analyzing complex survey data sage. University papers 71 Quantitative Applications in the Social Sciences, 1989.
22. Kish L, Frankel M. Inference from complex samples (with discussion). Journal of the Royal Statistical Society, Series B 1974; **36**: 1–37.
23. Lehtonen R, Pahkinen EJ. Practical methods for design and analysis of complex surveys. Chicester: John Wiley and Sons, 1965.
24. Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random-effects model: an example. J Roy Stat Soc Ser C, Appl Stat 2001; **50**: 325–335.
25. International Animal Health Code, 10th ed. OIE, Paris, France, 2001.

26. Van Wuijckhuise L, Bosch J, Franken P, Frankena K, Elbers ARW. Epidemiological characteristics of bovine herpesvirus 1 infections determined by bulk milk testing of all Dutch dairy herds. Vet Rec 1998; **142**: 181–184.
27. Thiry E, Saliki J, Schwers A, Pastoret PP. Parturition as a stimulus of IBR virus reactivation. Vet Rec 1985; **116**: 599–600.
28. Thiry E, Saliki J, Bublot M, Pastoret PP. Reactivation of infectious bovine rhinotracheitis virus by transport. Comp Immunol Microbiol Infect Dis 1987; **10**: 59–63.
29. Engels M, Ackermann M. Pathogenesis of ruminant herpesvirus infections. Vet Microbiol 1996; **53**: 3–15.
30. Greiner M, Gardner IA. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev Vet Med 2000; **45**: 3–22.