ESTIMATING THE FORCE OF INFECTION FROM SEROLOGICAL DATA: THE IMPACT OF CLUSTERING, INCOMPLETE DATA AND DIAGNOSTIC UNCERTAINTY

Niel H., Faes, C., Bollaerts, K., Aerts, M., Shkedy, Z.

Center for Statistics
Universiteit Hasselt, Campus Diepenbeek
Agoralaan 1, gebouw D, 3590 Diepenbeek, Belgium

Geys H.

Johnson & Johnson PRD, Belgium

Mintiens K., Laevens H.

Veterinary and Agrochemical Research Centre, Belgium

Boelaert F.

European Food Safety Authority, Italy

ABSTRACT

Epidemiology deals with the study of infectious diseases and their determinants within a given natural population. Infectious diseases data are complicated and statistical modelling should account for these burdens. In this paper, the impact of (1) clustering, (2) missing data and (3) diagnostic uncertainty on the estimation of the force of infection of the bovine herpesvirus-1 in Belgian cattle is established and several methods to deal with these complications are proposed.

SAMENVATTING

Epidemiologie is de studie van infectieuze ziektes en hun determinanten in een gegeven natuurlijke populatie. Infectieziekte data zijn ingewikkeld en de verschillende moeilijkheden dienen die in rekening gebracht te worden bij het statistisch modelleren. In dit artikel, bestuderen we de impact van (1) clustering, (2) ontbrekende gegevens en (3) diagnostische onzekerheid op het schatten van de infectiedruk voor het bovine herpesvirus-1 in Belgisch vee. Er worden verschillende methoden voorgesteld om met deze moeilijkheden om te gaan.

1. INTRODUCTION

The seroprevalence survey of the bovine herpesvirus-1 (BoHV-1) in Belgian cattle [1], conducted in 1998, is a study of a transmissible disease in cattle, which is of economic importance and significance to international trade. A central characteristic of infectious disease dynamics is the transmission of the infection from infectious to susceptible subjects. The force of infection (FOI) is the rate of acquisition of the infection for a susceptible host. Empirical data show that, in general, the FOI is age-dependent. Like many other infectious diseases data, the BoHV-1 data suffer from several complications and thus statistical modelling has to deal with these.

A first complication is clustering. Indeed, animals within clusters (herds) have a higher chance of becoming infected once the infection is introduced into the herd. Thus, individual responses are more homogeneously distributed within herds than in the whole population. There exist several methods to deal with clustering [2].

A second complication is the not unlikely occurrence that some subjects have one or more missing values. If the missingness is ignorable as defined by [2], the analysis can be based on the so called complete cases, i.e. all
observations for which all values are observed. If, however the missingness is non-ignorable, analyses can be affected by merely using the complete cases. Several methods to handle missing data are known [3]. None of them are without limitations.

A third complication is the diagnostic uncertainty when modelling the seroprevalence (apparent prevalence) instead of the true prevalence. The true prevalence of a disease is the proportion of a given population that is affected with that disease, whereas the seroprevalence is the proportion of test-positives for that disease. Evidently, main interest lies in this true but unobserved prevalence rather than in the observed test-specific seroprevalence. However, often serological data are analyzed such that only conclusions with respect to the observed seroprevalence are justified.

2. MATERIAL AND METHODS

In the present dataset, from a Belgian 1998 sero-survey, the response variable is the gB-test result for the presence of antibodies to BoHV-1. Additionally, age and origin (purchased yes/no) of the cows were recorded. Unfortunately, there was a considerable amount of animals for which the origin was not recorded. The FOI as a function of age was derived from the sero-prevalence function. Let us now present the different methods to deal with the complications of clustering, missing data and diagnostic uncertainty, respectively.

There are several ways to deal with clustering, some of which estimate population-averaged measures of effect and some of which estimate herd-specific (cluster-specific) measures of effect. Generalized estimating equations (GEE) can be used to estimate population-averaged measures while accounting for the clustering. Herd-specific profiles can be studied by using a generalized linear mixed model (GLMM) [4]. A more detailed study can be found elsewhere [5].

One of the techniques to deal with missing values, which gained a lot of attention recently, is the use of weighted estimating equations [6], where each contribution of a case is weighted with the inverse of the probability that this case is observed. In this way cases with a low probability to be observed gain more influence in the analysis and thus represent the missing values. One can look at this approach as an implicit imputation of missing values. Dealing with both missing values and clustering can be done using a weighted GEE, where the weights are those inverse probability weights [7], which are preferably estimated nonparametrically.

The diagnostic uncertainty related to modelling the seroprevalence instead of the true prevalence is determined by two test characteristics, i.e., diagnostic sensitivity and specificity. True prevalence is then modelled using serological data and prior information on diagnostic sensitivity and specificity. We propose the use of Bayesian techniques to study the impact of this uncertainty [8].

3. RESULTS

To show the impact of clustering on the estimation of the seroprevalence, the estimated FOI curves, based on a logistic model (ignoring clustering), on GEE (population averaged model) and on a random intercepts model (herd-specific), are shown in the next figures. The logistic regression model clearly underestimates the variability. The random effects model shows the large differences between herds.
To illustrate the effect of ignoring missing data, Figure 2 describes the results of modelling the FOI as a function of age, based on all cows, termed as ‘all cases’ (AC) on the one hand, and on the other hand based on those cows for which origin is observed, termed as ‘complete cases’ (CC). Next to showing the effect of merely using the complete cases, the use of weighted estimating equations (WCC) is illustrated to correct the CC-analysis.

Figure 2: The force of infection from a population averaged point of view as a function of age based on ‘all cases’ (AC), on the ‘complete cases’ (CC) and using weighted generalized estimating equations (WCC).

The effect of test misclassification is illustrated in the next two graphs that compare a model assuming a perfect test (Se = Sp = 1) with a model accounting for test misclassification with sensitivity and specificity as high as 0.995 and 0.990, respectively. Although virtually no differences are observed with respect to the estimated prevalence (left panel), larger differences are observed on the scale of the force of infection (right panel).

Figure 3: The prevalence (left panel) and the force of infection (right panel) as a function of age for a model assuming perfect test (dashed lines) and a model accounting for test misclassification (full lines).

### 4. DISCUSSION

In this paper, it is shown that specific models are needed to estimate the force of infection while dealing with clustering and missing values, both complications epidemiological data often suffer from. Moreover, main interest lies in the true but unobserved prevalence rather than in the observed test-specific seroprevalence. However, often serological data, as in this case, are analyzed such that only conclusions with respect to the observed seroprevalence can be drawn.

To deal with correlated data both generalized estimating equations and generalized linear mixed models have been proposed while inverse probability weights have been shown to be able to handle missing values. Bayesian methodology was used to assess the impact of diagnostic uncertainty. The proposed methods deal with two
additional complications, the constraint that the force of infection has to be positive and an informative cluster size, for which the details were omitted from this paper.

In practice, policy makers base their decisions on conducted surveys. Therefore, correct modelling of the epidemiological quantities of interest and correct interpretation of model parameters are crucial.

5. ACKNOWLEDGEMENTS

We further gratefully acknowledge support from the Fund of Scientific Research (FWO grant G.0393.04) Flanders, Belgium, from the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium and from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

REFERENCES