Preventive Veterinary Medicine 112 (2013) 348–354

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Adjusting for multiple clinical observers in an unbalanced study design using latent class models of true within-herd lameness prevalence in Danish dairy herds

N.D. Otten[a,*], N. Toft[a], H. Houe[a], P.T. Thomsen[b], J.T. Sørensen[b]

[a] Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark
[b] Department of Animal Science, Aarhus University, Blichers Allé 20, DK-8830 Tjele, Denmark

ARTICLE INFO

Article history:
Received 5 October 2012
Received in revised form 19 August 2013
Accepted 2 September 2013

Keywords:
Observer effect
Misclassification
Prevalence
Cattle
Lameness
Grazing

ABSTRACT

The elimination of misclassification bias introduced by multiple observers was evaluated and discussed based on an illustrative example using lameness prevalence in 80 Danish dairy herds. Data from 5073 cows from loose-housed cubicle herds larger than 100 cows were included in the analysis. Four trained observers performed clinical scoring on cow level and undertook a calibration test with 39 video sequences. The calibration test served both the purpose of estimating inter-observer agreement (PABAK = 0.69) in accordance with previous results and to estimate the sensitivity (Se) and specificity (Sp) for each observer. In the absence of a gold standard for the clinical observations, a latent class analysis (LCA) evaluating the true within-herd lameness prevalence was used. Sensitivity amongst observers was fairly low (0.24–0.81) inducing a general underestimation of the true prevalence. Comparative analyses were made to assess the effect of grazing on the lameness prevalence in order to demonstrate the consequences of using unadjusted apparent prevalences (AP) compared to the true prevalences (TP). Lameness prevalence was higher in grazing herds using AP estimates (19.0% zero-grazing, 20.2% grazing); while the TP estimates showed the expected higher lameness prevalence in zero-grazing herds (42.3% vs. 35.9%). Hence, this study emphasizes the importance of adjusting for observer Se and Sp to obtain true prevalence and avoid false interpretation.

1. Introduction

On-farm assessment of health and welfare on animal level requires the use of clinical and behavioural observations on individual animals. It is well known that such data are prone to misclassification and that differences between individual observers occur (Baadsgaard and Jørgensen, 2003). To alleviate the observer effects, training and calibration is seen as an essential part of studies involving multiple observers. Evaluation of the training and calibration of observers before, during and after study completion can be assessed as inter-observer agreement (IOA). Several studies rely only on kappa or prevalence adjusted bias adjusted kappa (PABAK) values as measures of agreement validating the given clinical condition measured. As an example, the clinical measures and assessment schemes used within the global welfare assessment protocol Welfare Quality® (WQ) were partly selected based on their validity in terms of inter-observer agreement. Within the overall WQ assessment of the clinical measures, the IOA for e.g. lameness was evaluated at four successive training sessions, improving the mean PABAK values from 0.6 to 0.7 over time (Brenninkmeyer et al., 2007), which were considered as sufficient levels of agreement. Kappa and PABAK

* Corresponding author. Tel.: +45 35 33 15 20; fax: +45 35 33 30 22.
E-mail address: nio@sund.ku.dk (N.D. Otten).

0167-5877/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.prevetmed.2013.09.002

hold one major disadvantage: they only yield information about the agreement and cannot tell whether disagreement between observers is systematic. Hence, they do not eliminate the problem of observer subjectivity. Another challenge concerning the reliability of the included animal-based measures is that no real consensus on the limits for discriminating between acceptable and unacceptable agreement exists, although several limits are proposed in the literature (Knierim and Winckler, 2009).

Consider as an example a study aimed at investigating the effect of grazing on the occurrence of lameness at herd level. This study could be performed using a number of herds with and without grazing in which a sample of cows were scored according to their lameness status (lame/not lame). In order for such a study to be conducted in a reasonably short time span, multiple observers each visiting separate herds are needed. Hence, data from such a study will consist of observed (or apparent) prevalences for a number of grazing and non-grazing herds. The problem with apparent prevalences is that they do not easily compare across populations as they represent the joint effect of observer bias and the true underlying lameness prevalence, thus resulting in potentially biased and misleading results and conclusions. In particular if the observers are not randomly allocated to the grazing and non-grazing herds, confounding bias is likely to occur. However, the true lameness prevalence may be derived from the observed lameness prevalence given that information about the sensitivity (Se, probability that a truly lame cow is classified as lame) and specificity (Sp, probability that a truly not lame cow is classified as not lame) is available for each observer. Unfortunately, getting reliable estimates of properties of the tests involved, i.e. in our case, estimates of Se and Sp for the individual observers of lameness remains a challenge. While literature and pilot-studies might be relevant for Se and Sp of e.g. serological, bacteriological or histopathological tests, the notion that a diagnostic test/mechanism ideally must be evaluated in the population where it is intended to be used, holds probably even stronger for clinical or observational data (Greiner and Gardner, 2000). Furthermore, the true underlying condition for most welfare related issues is generally unobservable. This requires models which do not rely on a perfect test for comparison. Latent class models (Hui and Walter, 1980) provide a tool for estimating Se and Sp of diagnostic tests in the absence of a perfect reference test given certain assumptions about the tests and the test subjects. Hence, models which allow for adjustment of the essentially unknown misclassification caused by different observers should be utilized to obtain reliable and unbiased results of general clinical data regarding health and welfare of animals.

Lameness is an animal welfare problem and almost 40% of all Danish dairy cows are lame to some degree (Jørgensen et al., 2010). Furthermore, Thomsen et al. (2004) reported locomotor disorders being the reason for euthanasia in 40% of the euthanized dairy cows in Denmark. Numerous risk factor studies have been performed in the past with different lameness scoring systems being used for prevalence estimation, but without stating observer Se and Sp.

The overall objective of this study was to provide and discuss a framework for an unbalanced design with multiple observers using latent class models to estimate the true prevalence, illustrated and motivated by the example of the effect of grazing on lameness at herd level.

2. Materials and methods

2.1. Study design and model

To meet the challenges outlined above, the framework that we propose must have two distinct features. Firstly, comparison between groups must be done using the true prevalence in order to adjust for the misclassification imposed by the use of different observers. Secondly, data must allow inference about the Se and Sp of the observers. The procedures and dataflow involved in the modelling are summarized in Fig. 1.

Inference about Se and Sp of the observers can be achieved in several ways. The predominant problem is the lack of a perfect reference observer or so called gold standard. As stated above, this problem can be circumvented by applying latent class models (Hui and Walter, 1980). The general assumptions underlying these models are sometimes referred to as the Hui–Walter paradigm: (I) the data must be from two or more populations with differing prevalence; (II) the Se and Sp of the two or more individual observers must be constant across these populations; (III) the observers are considered to be conditionally independent given the underlying disease/welfare condition. However, in the original paper it is noted that also 3 or more conditionally independent observers and just one population will yield sufficient information (i.e. degrees of freedom) to allow estimation of the required parameters. In our example, we will make use of this setup, but we will address more general designs in the discussion.

For the example that will be elaborated further in the next section, data from a reliability study were available to calculate Se and Sp of the 4 observers, using the following Bayesian model:

$$\mathrm{O}_{ij} \sim \mathrm{Bernoulli}(\mathrm{OP}_{ij}), \quad i = 1, 2, 3, 4,\ j = 1, \ldots, 39$$

$$\mathrm{OP}_{ij} = \mathrm{Se}_i \times \mathrm{TC}_j + (1 - \mathrm{Sp}_i) \times (1 - \mathrm{TC}_j), \quad i = 1, 2, 3, 4,\ j = 1, \ldots, 39$$

$$\mathrm{TC}_j \sim \mathrm{Bernoulli}(P), \quad j = 1, \ldots, 39$$

$$\mathrm{Se}_i \sim \mathrm{Beta}(1, 1), \quad i = 1, 2, 3, 4$$

$$\mathrm{Sp}_i \sim \mathrm{Beta}(1, 1), \quad i = 1, 2, 3, 4$$

$$P \sim \mathrm{Beta}(1, 1)$$

where O_{ij} is the observation of the ith observer on the jth cow (in the reliability study), this observation follows a Bernoulli distribution, i.e. is either 0 (not lame) or 1 (lame); OP_{ij} is probability of observing a 1; TC_j is the true condition of the jth cow, which is following a Bernoulli distribution, with probability P; Se_i and Sp_i are the Se and Sp of the ith observer, respectively. As we do not have any useful prior