Strategy for statistical modelling in analytic epidemiology: application to herd-level risk factors for bovine mastitis.

N. Bareille, C. Fourichon*, F. Beaudeau, H. Seegers H. Unit of Animal Health Management, Veterinary School - INRA, BP 40706, 44307 Nantes cedex 03, France

Summary

Association of 54 explicative variables with the incidence density of clinical mastitis (n = 237 herds) was investigated by multivariable Poisson regression. This paper aims at assessing consequences of different modelling strategies in a context of numerous explicative variables and possible reverse causal relationships. The results suggest that a modelling strategy without any preselection step and which accounts for intra- or inter-cluster relationships between variables should be preferred when applicable. Otherwise, cluster preselection should be used rather than a univariable preselection. In addition, an initial step of elimination of the reverse causal explicative variables should be performed to prevent from masking risk factors.

Introduction

Risk-factor analysis of multifactorial diseases at herd level implies to evaluate simultaneously a large number of risk factors. Multivariable statistical models can be used for this purpose. A preselection step of a set of putative risk factors is generally performed in the modelling process. Modelling strategies combining univariable and/or multivariable preselection steps are commonly implemented without a clear justification of the chosen approach [1, 3, 5, 6, 7]. The problem is that these strategies can lead to different sets of preselected variables and hence to finally identify different risk factors. This is a consequence of unexpected associations between independent variable [4]. A second problem arises when analysing data from new surveys aiming at updating risk analysis of well-known health disorders like mastitis. In herds with a high frequency of mastitis, farmers are advised to implement recommended herd management practices. Then, in cross-sectional surveys, statistical modelling may lead to evidence reverse causal relationships. The investigator may choose to remove or to keep these reverse causal independent variables in the modelling process. This paper aims at assessing consequences of different modelling strategies on the results (identified risk factors and their quantitative effect) in a context of numerous explicative variables and possible reverse causal relationships.

Materials and methods

Data were issued from a survey in 237 dairy farms located in west of France. 54 putative risk factors were selected and classified into 5 groups of management practices (herds and cows characteristics, milking procedures, housing conditions, management of dry cows, nutrition). Their association with the incidence density of clinical mastitis was investigated by multivariable Poisson regression with a backward selection process. 6 alternative modelling strategies were defined depending, first, on the elimination of the reverse causal explicative variables (ERCEV) or not, second, on a one- or two-step variable selection and, third, when a preselection step existed, on an univariable or multivariable (within each cluster i.e. groups of management practices variables) regression modelling (Table 1). In the preselection step, variables at a *P*-value < 0.30 were considered for further analysis. In the final step, retained explicative factors had a *P*-value < 0.10.

Table 1. Definition of the 6 alternative modelling strategies.

Modelling strategy	1SK	2UK	2MK	1SE	2UE	2ME
Elimination of the reverse causal explicative variables	No	No	No	Yes	Yes	Yes
Preselection of variable offered to the model	No	Yes	Yes	No	Yes	Yes
Method for preselection	-	Univ[1]	Multi[2] (cluster)	-	Univ	Multi (cluster)

[1] univariable; [2] multivariable : Poisson regression with all variables in a cluster of management practices.

Results

Two strategies led to the same final model: the one-step and the multivariable preselection strategies with ERCEV. The elimination of the reverse causal explicative variables led to a decrease of the number of explicative variables in the preselection and final steps whatever the preselection strategy (Table 2). The five reverse causal explicative variables identified described the milking machine and the milking procedures. In the strategy with a univariable preselection with ERCEV, 4 explicative variables were excluded in comparison to the other strategies with ERCEV leading to a lower R² (0.48 *vs*. 0.55): one explicative variable was excluded at the preselection step (*P*-value = 0.47 before backward elimination) and the three other were excluded at further steps (0.10<*P*-value<0.15). Only 6 real risk factors were common to the 6 strategies and displayed rate ratio varying from 1.16 to 1.43. For a given risk factor, the rate ratios estimated in the 6 strategies were almost constant.

Table 1. Number of explicative variables retained at the end of each step of the modelling process and coefficient of determination of the final model according to the modelling strategy.

Modelling strategy [1]	1SK	2UK	2MK	1SE	2UE	2ME
Nb variables in the preselection step	-	28	32	-	20	20
Nb variables in the final step	12	11	12	11	7	11
Number of RCEV[2]	4	3	2	0	0	0
Number of risk factors	8	8	10	11	7	11
Coefficient of determination	0.549	0.551	0.565	0.532	0.476	0.532

[1] see table 1; [2] reverse causal explicative variables.

Discussion

This paper did not aim at comparing all the possible modelling strategies in analytic epidemiology, but at assessing the consequences of common strategies.

The first step of a modelling strategy is often an univariable preselection step with a large variation in *P*-value to be considered for further analysis (0.10 [7]; 0.15 [1]; 0.20 [6]; 0.25 [4]). Our results suggest that these *P*-values are too low at this step. Indeed, the univariable preselection modelling led to exclude an explicative variable at an initial *P*-value = 0.47; its effect was probably masked by other variables included in the multivariable models. The one-step strategy which accounts for all possible existing relationships between variables (intra- or inter-cluster) should be preferred when applicable (available computer resources). Otherwise, when there are too many variables, a cluster preselection should be used rather than a univariable one. Clusters should then be defined according to possible structural partial multicollinearity [2].

An additional difficulty arises when risk factors of a health disorder have been identified for years and widely disseminated by farm advisors. In our study, this may be the case with the separate milking of mastitic cows. A significant relationship, reverse to biological assumptions is then evidenced (high incidence density when a separate milking is implemented). In such a case, the reverse variable is likely to influence the results of other variables. Therefore, it is preferable to exclude it from the model as soon as the reverse causal relationship is evidenced. When such a relation was not evidenced in the first step of the modelling process, we decided to rerun all the modelling process from start without this reverse variable.

In cross sectional surveys, to prevent from masking risk factors, we recommend that an initial multivariable step of elimination of the reverse causal explicative variables is implemented.

References

1 Bartlett, P.C., Miller, G.Y., Lance, S.E., Heider, L.E., 1992. Environmental and managerial determinants of somatic cell counts and clinical mastitic incidence in Ohio dairy herds. Prev. Vet. Med. 14, 195-207.

2 Dohoo, I.R., Ducrot, C., Fourichon, C., Donald, A., Hurnik, D., 1996. An overview of techniques for dealing with large numbers of independant variables in epidemiologic studies. Prev. Vet. Med. 29, 221-239.

3 Elbers, A.R.W., Miltenburg, J.D., De Lange, D., Crauwels, A.P.P., Barkema, H.W., Schukken, Y.H., 1998. Risk factors for clinical mastitis in a random sample of dairy herds from the southern part of the Netherlands. J. Dairy Sci. 81, 420-426.

4 Hosmer, D.W., Lemeshow, S., 1989. Applied logistic regression. John Wiley, New York, 307 pp.

5 Khaitsa, M.L., Wittum, T.E., Smith, K.L., Herderson, J.L., Hoblet, K.H., 2000. Herd characteristics and management practices associated with bulk-tank somatic cell counts in herds in official dairy herd improvement association programs in Ohio. Am. J. Vet. Res. 61, 1092-1098.

6 Peeler, E.J., Green, M.J., Fitzpatrick, J.L., Morgan, K.L., Green, L.E., 2000. Risk factors associated with clinical mastitis in low somatic cell count British dairy herds. J. Dairy Sci. 83, 2464-2472.

7 Schukken, Y.H., Grommers, F.J., Van de Geer, D., Erb, H.N., Brand, A., 1990. Risk factors for clinical mastitis in herds with a low bulk milk somatic cell count I. Data and risk factors for all cases. J. Dairy Sci. 73, 3463-3471.