Advantages and inconveniences of the Cox model compared with the logistic model: application to a study of risk factors of nursing cow infertility

F Bugnard, C Ducrot, D Calavas

Vet Res (1994) 25, 134-139

Advantages and inconveniences of the Cox model compared with the logistic model: application to a study of risk factors of nursing cow infertility

F Bugnard *, C Ducrot, D Calavas

Centre d'Écopathologie Animale, 26, rue de la Baïsse, 69100 Villeurbanne, France

Summary — The survival Cox model and the logistic model were compared on a data set obtained from an ecopathological survey relative to the risk factors of nursing cow infertility. The risk factors resulting from the 2 models were the same. The Cox model has the advantage of preserving the variable in its original quantitative form, and of using a maximum of information. However, very restrictive conditions of application of this model make its use rather limited.

ecopathology / survival model / logistic model

Résumé — Avantages et inconvénients du modèle de Cox par rapport au modèle logistique : application à l'étude des facteurs de risque de l'infécondité des vaches allaitantes. *Le modèle de survie de Cox et le modèle logistique ont été comparés sur un jeu de données issu d'une enquête d'écopathologie relative aux facteurs de risque de l'infécondité des vaches allaitantes. Les facteurs de risque issus des 2 modèles étaient les mêmes. Le modèle de Cox présente l'avantage de conserver à la variable étudiée sa forme quantitative d'origine et d'utiliser le maximum d'information. Cependant, les conditions d'application très restrictives de ce modèle sont une limite à son utilisation.*

écopathologie / modèle de survie / modèle logistique

INTRODUCTION

Logistic regression is widely used for investigation of risk factors in epidemiology in general and in the Centre d'Écopathologie Animale in particular. There are 2 main reasons for this. Firstly, the pathologies studied are often characterized as absent or present and we wish to explain the risk of an occurrence of the disease. In this sense, the logistic model is appropriate as it allows the prediction of the probability of an event. On the other hand, an advantage of the logistic model is that the parameters can be interpreted as being the logarithms of the odds ratios of explanatory variables. However, when the pathology studied is characterized by a time interval (*eg,* calving intervaland breeding-to-conception period), logistic regression seems less suitable. Indeed, this method requires a division of the period

* Correspondence and reprints

of time into discrete classes, which leads to an important loss of information and poses a problem of which group end points to choose. The survival models and, in particular the Cox model, appear, therefore, to be better adapted. These models, initially created for cancer research in order to study the survival time of patients after therapy, allow the study of time intervals without division into classes. In agreement with the logistic model, the parameters of the Cox model can be easily interpreted since they are the logarithms of the relative risks of explanatory variables. To compare their results, both the logistic and the Cox model were applied to the same data set collected in a survey carried out by the Centre d'Écopathologie Animale in order to highlight the risk factors of nursing cow infertility.

MATERIALS AND METHODS

The data were collected during a survey carried out from 1987 to 1989 by the Centre d'Écopathologie Animale in 116 farms in the Rhônes-Alpes and Centre regions and in the district of Yonne (Ducrot, 1993). In the course of this period, 3 583 mating cows were individually monitored after their first calving during the survey. The infertility of cows was determined by the calving interval. The covariates introduced in the logistic and the Cox model were selected from the hypotheses of the risk factors with the help of univariate analyses (χ^2, Logrank and Wilcoxon test) and only those with a 20% significant link with infertility were kept. The 20% level was chosen in order not to eliminate covariates which had little effect on the disorder when studied separately, but, if associated, might have a strong influence upon the disorder (Hosmer and Lemeshow, 1989). The selected factors were: the characteristics of the cow and the calf (parity, breed of the cow, number, sex, weight, and presentation of calves); the calving conditions (calving difficulty, characterized by the type of intervention, retained placenta, characterized by the fact that the placenta is not eliminated spontaneously within 24 h of calving, and acute metritis with general symptoms appearing within 2 d of calving); and the mating conditions (calving to turning-out-to-graze period, whether or not the cows turned out to graze before mating, cleanliness score of the cows at the time of turning them out to graze measured on 4 anatomic zones of the hind quarters (Faye and Barnouin, 1985), type of basal ration after calving, quantity of concentrates distributed after calving, fattening score at the time of turning out to graze determined by manual handling (Agabriel *et al*, 1986), fattening variation during winter). Two additional data elements were taken into consideration as adjusting factors: the geographic region and the type of winter housing.

In the following, the vector $z_j = (z_{j\,1}, z_{j\,2}, \ldots, z_{jp})$ refers to the p covariates measured for the individual j ($j = 1, \ldots, n$).

The logistic model

The calving interval was divided into 2 discrete classes; the cows were considered infertile if the calving interval was more than 368 d, or if they were not fertilized, or if they were fertilized belatedly, *ie* if the zootechnical objective of one calf per year was not reached. Thus, we tried to explain an event Y coded as present or absent (infertile/fertile).

If $P(Z)$ is the probability of a cow being infertile knowing $Z = (z_1, z_2, \ldots, z_p)$, the logistic model is defined as follows:

$$P(Z) = \frac{\exp\,(b_0 + b_1 z_1 + b_2 z_2 + \ldots + b_p z_p)}{1 + \exp(b_0 + b_1 z_1 + b_2 z_2 + \ldots + b_p z_p)}$$

where $b = (b_0, b_1, \ldots, b_p)$ is the vector of the model parameters.

The logistic model was adjusted to the data using the maximum likelihood estimation method. The resulting model was analyzed by comparing globally all estimated parameters b_i to 0, using the likelihood ratio test, and also by comparing each parameter to 0, using the Wald test (Hosmer and Lemeshow, 1989). The parameters not significantly different from 0 at the 5% level were eliminated. Finally, the goodness of fit of the final model was determined by the χ^2 test of Hosmer and Lesmeshow (1989) which compares the number of the predicted infertility cases to the number of the observed cases. This analysis was carried out with the SAS software, LOGISTIC procedure (SAS Institute Inc, 1990).

The Cox model

The calving interval was studied without being divided into classes. The studied interval was from calving to the end of the survey period for the cows that did not have a second calving before the end of the survey (censored cows).

Let the hazard function $h(t,Z)$ be the probability of calving at the time t for a cow knowing $Z = (z_1, z_2,..., z_p)$. The Cox model is:

$$h(t,Z) = h_0(t) \exp (b'_1 z_1 + b'_2 z_2 + ... + b'_p z_p)$$

where $b' = (b'_1, b'_2, ..., b'_p)$ is the vector of the model parameters. These parameters are estimated by the maximum likelihood method using the Cox partial likelihood; $h_0(t)$ remains an unspecified function which does not need to be estimated (Cox and Oakes, 1984).

The analysis was performed in 2 stages. The utilization of the Cox model is bound to the verification of the proportional hazards assumption (which means, in other words, that the effect of the factors must be independent of time), and so the first stage consisted in verifying this hypothesis by representing the logarithm of the negative logarithm of the survival function of each class as a function of time for each covariate (fig 1). According to the proportional hazards assumption the curves must not cross each other (Lee *et al*, 1989). Covariates presenting many classes did not verify this hypothesis, and so we grouped classes in order to obtain variables satisfying this application condition. For the cleanliness score, the curves of the 2 classes did cross each other (fig 2). Since grouping was impossible, this variable was integrated in the model but the estimated parameters could not be interpreted. In the course of a second stage, the Cox model was applied to the whole data set. As for the logistic model, the estimated parameters were analyzed by the likelihood ratio test and the Wald test (Cox and Oakes, 1984). The covariates were eliminated from the model in the same way as in the logistic model (parameters not significantly different from 0 at the 5% level). This analysis was carried out with the SAS software, PHREG procedure (SAS Institute Inc, 1991).

The methods that allow an estimation of the goodness of fit of the Cox model have not yet been completely developed. They are based mainly on graphical methods and were not implemented in this study.

RESULTS

The 2 final models gave the same risk factors (table I). The same classes were sig-

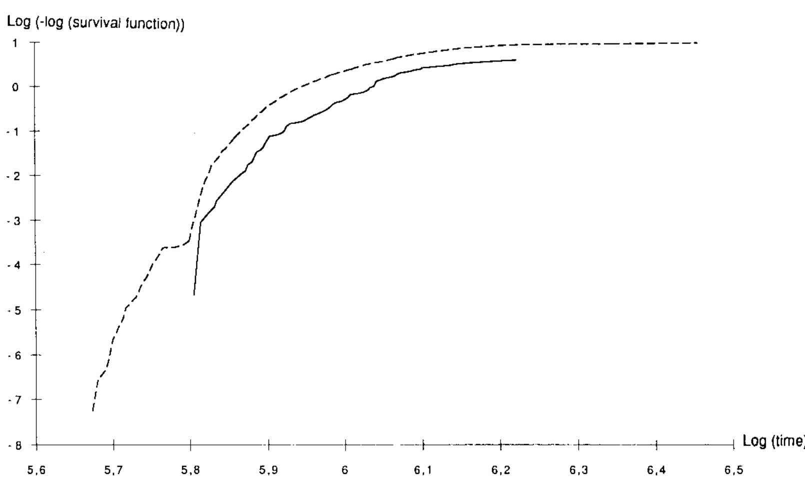

Fig 1. Logarithm of the negative logarithm of survival function *versus* logarithm of time for no retained placenta (– – –) and for retained placenta (—).

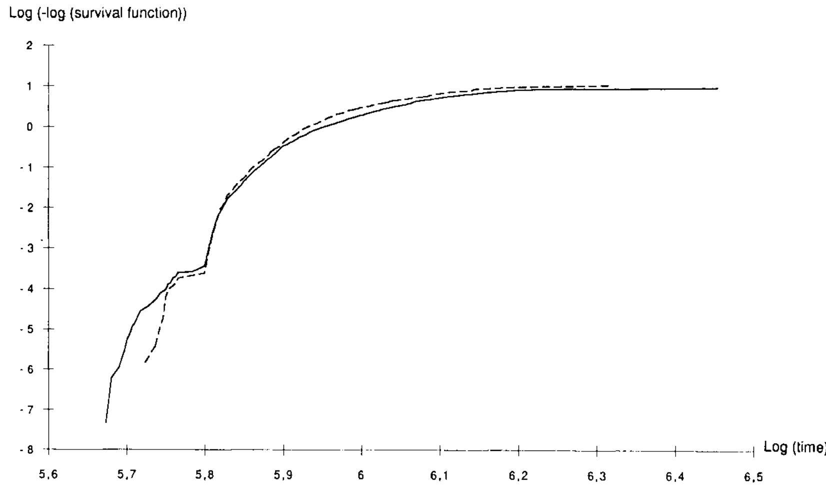

Fig 2. Logarithm of the negative logarithm of survival function *versus* logarithm of time for cleanliness score ≤ 2 (– – –) and for cleanliness score > 2 (—).

nificant for parity, acute metritis, and cleanliness score. For retained placenta, the 2 models gave almost the same P values (logistic model: $P = 0.04$; Cox model: $P = 0.0521$).

However, there are 2 differences. In the logistic model, for the variable 'calving difficulty', the parameters of the classes 'assistance with calf puller (1 person)' and 'forcible extraction' were not significantly different from 0, but they were in the Cox model. On the contrary, for the variable 'calving to turning-out-to-graze period' the parameters of the classes '< 1 month', '1–2 months' and '2–3 months' were not significantly different from 0 in the Cox model, but they were in the logistic model.

Finally, for the 2 adjusting variables, it was observed that: (i) for the type of winter housing, the same results were obtained in the 2 models; and (ii) for the geographic region, the classes 'Loire' and 'Centre/Yonne' were positioned in the same way relative to 'Rhône-Alpes' (Loire being more favorable, 'Centre/Yonne' less favorable). However, 'Loire' was significant only for the logistic model, while 'Centre/Yonne' was significant only for the Cox model.

DISCUSSION

Considering risk factors, the results of both models were equivalent. In our example, the loss of information due to the division into classes of the calving interval did not seem to influence the results.

The Cox model has 3 main advantages. First, it allows us to take into account all the cows even if they were not observed for the same period of time and even if the studied event (next calving) did not take place before the end of the survey. This advantage was arbitrarily blurred to a certain extent in our example insofar as the cows that did not have a second calving could be integrated in the logistic model (as they were declared infertile).

Second, it does not require the division of the studied variable into discrete classes. This is an advantage as it allows us to keep the information as precise as possible and

Table I. Logistic and Cox model results (parameter estimates, 95% confidence interval and *P* value for Wald test).

Factors	*Logistic model odds ratio* *Estimates*	*95% CI*		*Groups* *P*	*Cox model relative risk*	*Estimates*	*95% CI*		*P*
Parity	3.1	2.3	4.2	0.0001	1	0.67	0.588	0.763	0.0001
	1				2–3–4				
	1.2	0.98	1.5	0.07	≥ 5	1			
Calving difficulty	1				No assistance				
	1.2	0.9	1.6	0.09	Easy assistance	1			
	1.2	0.9	1.6	0.15	Assistance with calf				
					puller (1 person)	0.848	0.752	0.957	0.0074
	1.5	0.9	2.3	0.07	Forcible extraction				
	5.3	1.7	16.4	0.003	Caesarean	0.45	0.284	0.713	0.0007
Retained placenta	1				No	1			
	1.8	1.0	3.2	0.04	Yes	0.749	0.56	1.003	0.052
Acute metritis	1				No	1			
	4.2	1.3	12.9	0.01	Yes	1.449	0.288	0.70	0.0004
Calving to turn out to graze	1				Calving on pasture	1			
	1.3	0.98	1.9	0.05	< 1 month				
	1.5	1.07	2.1	0.01	1–2 months	0.904	0.784	1.042	0.1648
	2.0	1.4	2.9	0.0002	2–3 months				
	2.5	1.7	3.6	0.0001	≥ 3 months	0.785	0.659	0.935	0.0067
Cleanliness score	1				≤ 2	1			
	1.3	1.0	1.6	0.02	> 2	0.861	0.773	0.959	0.0064
Geograpic region	1				Rhône-Alpes without Loire	1			
	0.7	0.5	1.0	0.05	Loire	1.1	0.962	1.257	0.163
	1.2	0.9	1.5	0.2	Centre and Yonne	0.878	0.777	0.991	0.0356
Housing	1				Loose	1			
	1.4	1.1	1.9	0.009	Tethered	0.843	0.734	0.96	0.0155
	1.1	0.8	1.4	0.3	Both	1.025	0.906	1.16	0.6916
	1.2	0.8	1.9	0.3	No	1.015	0.815	1.265	0.8929

eliminates the always difficult problem, of which group end points to choose.

Finally, it allows us to formulate the results according to the increase in the studied time interval in d when the factor is present (Lee *et al,* 1989). For example, acute metritis increases the calving interval by 23 d. For popularization, this way of presenting the results is more comprehensible than the presentation in terms of the odds ratio or even of relative risk.

Despite all these advantages the use of the Cox model is very limited by its very restrictive application conditions (the hazards should be proportional). Indeed, we noticed in our example that if the studied variable consisted of a large number of classes, the hazards were rarely proportional all along the curve. This leads to group classes. In extreme cases, this condition can prevent us from integrating certain variables into the model.

However, problems only arise if the deviation from the proportional hazards assumption is large and the frequency of the studied event is high (the second calving, in our case). In this case estimated parameters are greatly influenced (Ingram and Kleinman, 1989) and therefore cannot be interpreted.

As for the power of the 2 models, the Cox model and the logistic model seem to be statistically equivalent for identifying the risk factors when the percentage of the studied events is low, whereas the Cox model is more powerful when the percentage is high (Annesi *et al,* 1989). The conclusion is that when the follow-up time is sufficiently short or the survival rate is high, the choice of model must depend on convenience (for the user), availability of computer software, and expertise.

REFERENCES

Agabriel J, Giraud JM, Petit M (1986) Détermination et utilisation de la note d'état d'engraissement en élevage allaitant. *Bull Techn CRZV Theix INRA* 66, 43-50

Annesi I, Moreau T, Lellouch J (1989) Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Stat Med* 8, 1515-1521

Cox DR, Oakes D (1984) *Analysis of Survival Data.* Chapman and Hall, London, 201 p

Ducrot C (1983) Approche biométrique des facteurs de risque des pathologies d'élevage à partir des enquêtes d'écopathologie. Application à l'infécondité des vaches allaitantes. Université Claude Bernard, Lyon, Thèse de doctorat, 118 p

Faye B, Barnouin J (1985) Objectivation de la propreté des vaches laitières et des stabulations – L'indice de propreté. *Bull Techn CRZV Theix INRA* 59, 61-67

Hosmer DW, Lemeshow S (1989) *Applied Logistic Regression.* Wiley, New York, 307 p

Ingram DD, Kleinman JC (1989) Empirical comparisons of proportional hazards and logistic regression model. *Stat Med* 8, 525-538

Lee LA, Ferguson JD, Galligan DT (1989) Effect of disease on days open assessed by survival analysis. *J Dairy Sci* 72, 1020-1026

SAS Institute Inc (1990) *SAS/STAT user's guide 2* Version 6, Gary, NC, 4th edition

SAS Institute Inc (1991) *SAS Technical Report SAS/STAT Software*: the PHREG Procedure. Version 6, Gary, NC, 1st edition