

Tree-Based Methods as an Alternative to Logistic Regression in Revealing Risk Factors of Crib-Biting in Horses

Krisztina Nagy, PhD, Jenő Reiczigel, PhD, Andrea Harnos, PhD, Anikó Schrott, MSc, and Péter Kabai, PhD

ABSTRACT

Determining the risk factors might help in designing prevention of crib-biting. Logistic regression is a commonly used statistical method for finding risk factors, but tree-based methods are also getting more popular. An important difference between these two statistical approaches is that logistic regression makes a number of assumptions about the underlying data, whereas tree-based methods do not. Another difference is that logistic regression can be used to derive odds ratios for the significant risk factors, whereas tree-based methods create a tree where the ramifications represent the risk factors. The probability of occurrence is assigned to each end of branch in the tree. Data of horses used for noncompetition purposes were analyzed with three statistical approaches: logistic regression, classification tree, and conditional inference tree methods. By this, we compared the advantages and disadvantages of these statistical methods. No difference was found between the two tree-based methods regarding the structure and prediction accuracy of the trees. Compared to them, logistic regression revealed fewer risk factors, and also the number of the stereotypic horses classified correctly by the model was less. The representation of the tree-based methods is closer to medical reasoning and also high-order interaction of the risk-factors can easily be visualized. Our results suggest that tree-based methods can be a new alternative in revealing risk factors, even if used alone or together with logistic regression.

Keywords: Crib-biting; Risk factors; Classification tree; Conditional inference tree; Logistic regression

INTRODUCTION

Crib-biting is the most common stereotypic behavior in horses, which involves grasping of a fixed object with the incisors and emitting a grunting sound called wind-sucking. Previous studies indicate that breed type, weaning process, housing and management conditions, and feeding regime have a strong effect on its development, and it is a commonly held belief that horses may learn to crib-bite from affected horses.¹⁻⁴ Crib-biting is very difficult to treat; therefore, efforts should be concentrated on prevention. Determining the risk factors is conducive to the design of protective measures.

Classification methods revealing risk factors of a certain disease provide not just a better understanding of the disease, but are also useful to classify individuals into risk groups with some certainty. Logistic regression is a commonly used statistical method for finding risk factors, which is a standard method for predicting a dichotomous dependent variable. A logistic regression model is a linear regression equation in which the response variable is the log odds.⁵ The tree-based methods, such as the classification and regression tree (CART) and the conditional inference tree analysis, use a form of binary recursive partitioning. If the outcome variable is measured on a continuous scale, the method is called regression tree, whereas in case of a categorical outcome variable (like crib-biting in the present study) it is called classification tree. Tree-based methods split the sample step by step into smaller and smaller groups according to a mathematical condition. There are several variants of tree-based methods with different splitting criteria. For example, one of the oldest tree classification methods, the CHAID (Chi-square Automatic Interaction Detector) technique uses an F test if the dependent variable is continuous and χ^2 if the variable is categorical to decide which group to split.⁶ Out of the several criterion functions, Gini-index is most often used.⁷ The basic idea is to consider all possible splits and choose the best predictor and the best split to maximize “purity,” that is, homogeneity of the child nodes. Variables might be selected repeatedly on different levels of the tree, also with different thresholds.⁸ After an initial large tree is built, pruning is done to remove any overfitting to the training

From the Szent István University, Faculty of Veterinary Sciences, Budapest, Hungary.
Reprint requests: Krisztina Nagy, PhD, Szent István University, Faculty of Veterinary Sciences, Budapest, István u. 2, H-1078, Hungary.
0737-0806/\$ - see front matter
© 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvevs.2009.11.005

data. An automatic pruning method is the cost-complexity pruning based on cross-validation. It reduces the number of branches of the tree (variables selected as risk factors) and minimizes the percentage misclassified at the same time. A plot of percentage misclassified against the number of terminal nodes helps determine the optimal tree-size.

Recursive fitting procedures have been reported to have two fundamental problems: overfitting and a selection bias toward covariates with many possible splits or missing values. Although pruning procedures are able to solve the overfitting problem, the variable selection bias still seriously affects the interpretability of tree-structured regression models. Conditional inference tree is a rather new tree-based method, which estimates a regression relationship by binary recursive partitioning in a conditional inference framework. Therefore, it selects variables in an unbiased way. A statistically motivated stopping criterion (e.g., c_{quad} -type test statistics) is used, and the partitions induced by this recursive partitioning algorithm are not affected by overfitting. Partitions obtained from conditional inference trees have been reported to be generally closer to the true data partition compared to partitions obtained from an exhaustive search procedure with pruning.⁹

Tree-based methods have been widely used in computer sciences, and it is also getting more popular in life sciences, such as in human health care,^{7,10,11} medical decision-making,¹² or psychiatry¹³, as well as in the field of ecology¹⁴ or animal behavior.¹⁵ However, application of tree-based methods in veterinary sciences is not so prevalent. In this study, we would like to show an example on how CART and conditional inference trees can be applied in predicting risk factors of stereotypic behavior in horses. To illustrate the effectiveness of tree-based methods and to compare them to logistic regression, a previously published data set on risk factors of crib-biting in horses¹⁶ was reanalyzed.

METHODS

Data

To detect potential risk factors of crib-biting, a questionnaire survey was carried out on 287 horses. The survey items focused on housing and management conditions, food regime, stereotypies, and problematic behavior performed by the individual horse or by a horse in its visual contact.¹⁶ Data of horses used for noncompetition purposes ($N = 126$) were analyzed by using three statistical approaches: logistic regression, CART, and conditional inference tree methods.

Classification Techniques

Logistic Regression. Variables were preselected by univariate logistic regression models (Generalized Linear Model, GLM). Variables with $P < .1$ were considered for initial

inclusion in the multivariate GLM. The model was built manually by using a backward elimination process; variables with a $P > .1$ were excluded. GLM was applied with binomial error distribution and logit link function. The exponentials of regression coefficients (β) in the final model were interpreted as odds ratios. The significant variables in the final logistic regression model represented the risk (odds ratio > 1) or preventive factors (odds ratio < 1). The best model was chosen by using the Akaike information criterion. Diagnostic plots of the residuals and standardized residuals were used to check the normality and variance homogeneity assumptions of the final model.

Classification Tree and Conditional Inference Tree. The brief description of the applied binary recursive partitioning algorithm, omitting the mathematical details, is as follows. The method starts with one single group, the whole sample. In the first step, an explanatory variable and a threshold is selected and the sample is split into two groups: one in which the value of the selected explanatory variable is over the selected threshold, and the other in which it is below the threshold. That variable and threshold is selected that leads to the split with the most “pure,” that is, most homogeneous groups with respect to the outcome variable. In each subsequent step a group, an explanatory variable and a threshold is selected, and the selected group is split into two, based on the selected variable and threshold. The criterion defining the homogeneity of groups, and so driving the whole partitioning process, is called the split criterion, or splitting function. There are several split criteria, of which we applied the most often used one, the so-called Gini index. The process results in a tree-like structure of groups, also called nodes, in which each node has two “child nodes.” Terminal nodes, also called branches of the tree, define the classification of subjects.

Classification tree was built by splitting each node until its child nodes contained less than three observations. Gini index was used as splitting function and 10-fold cross-validation was applied to evaluate performance of the classification. After the initial large tree had been constructed, pruning was performed to reduce the size of the tree.

Ten-fold cross-validation means that the following procedure is repeated 10 times: a 10% random sample is selected from the data, the model is fitted to the remaining 90%, and prediction is made for the selected 10% from the fitted model. Classification performance is calculated by pooling classification results of the 10 replications.

Conditional inference trees were constructed with c_{quad} -type test statistics and $\alpha = 0.10$, with and without simple Bonferroni correction. Each split needed to send at least 1% of the observations into each of the two child nodes.

To compare the three statistical approaches, we examined prediction accuracy via the indices of sensitivity (correctly classified stereotypic horses) and specificity

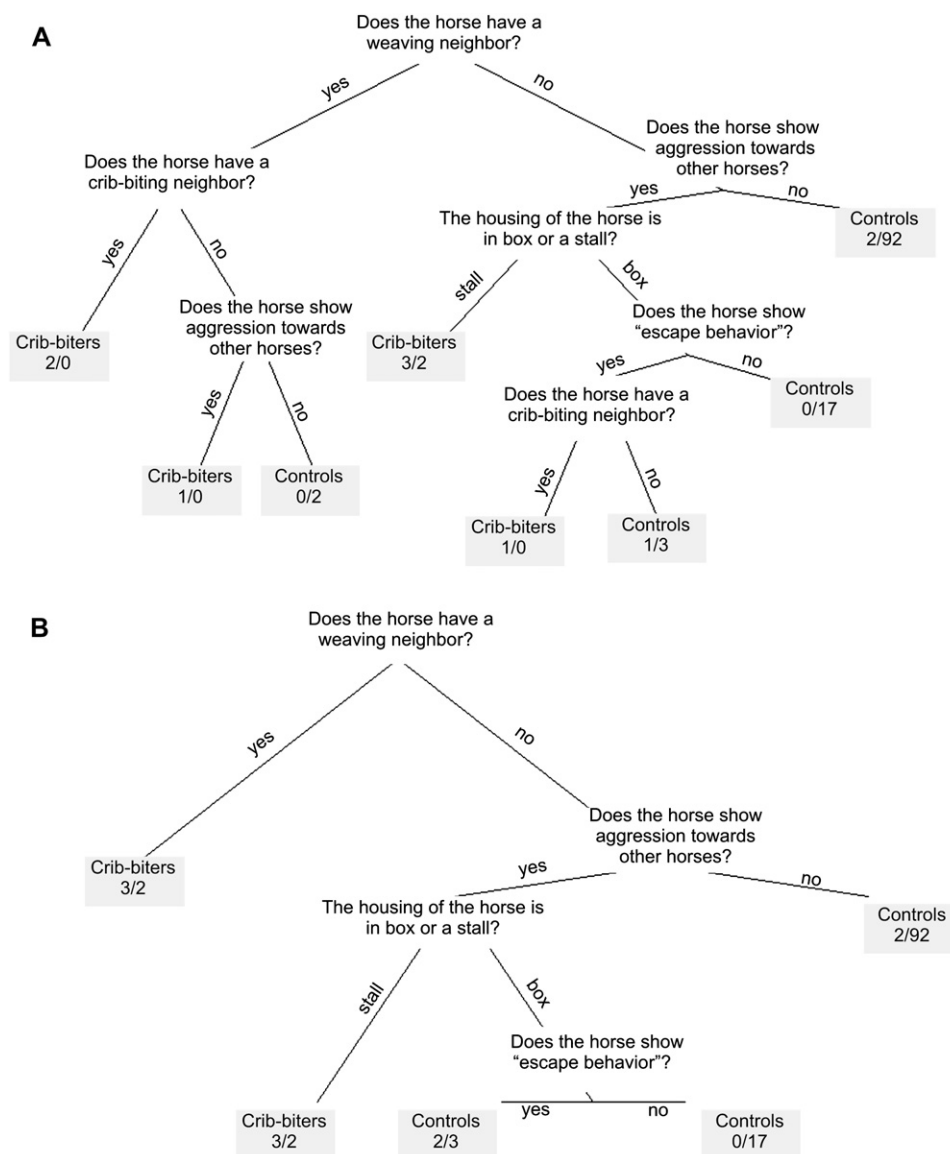


Figure 1. Classification tree before (A) and after (B) pruning. The ramifications of the tree represent the risk factors. Terminal nodes are categorized as crib-biters or controls (nonstereotypic) according to the number of crib-biting and control horses observed in the sample at that node (the numbers displayed at each terminal node).

(correctly classified nonstereotypic horses) of the models. All analyses were carried out using the R 2.7.2. Statistical Software.¹⁷ CART analysis and conditional inference trees are implemented in the rpart and party add-on packages to the R system for statistical computing.

RESULTS

Risk Factors

Logistic Regression. Eight variables were selected by the univariate GLM into the logistic regression model: presence of a weaving neighbor, presence of a crib-biting

neighbor, presence of a box-walking neighbor, presence of an aggressive neighbor, aggression toward horses, get out behavior (door or tier opening), box-walking, and wood-chewing. The final model contained only three variables. Horses that were kept in the neighborhood of another crib-biting or weaving horse were more likely to show crib-biting themselves, with odds being 7 and 21 times greater, respectively. Horses with aggressive behavior toward other horses had 11 times greater risk of performing crib-biting than nonaggressive horses. The logistic regression model classified horses without stereotypic behavior more accurately (specificity: 99%) than crib-biting horses (sensitivity: 40%).

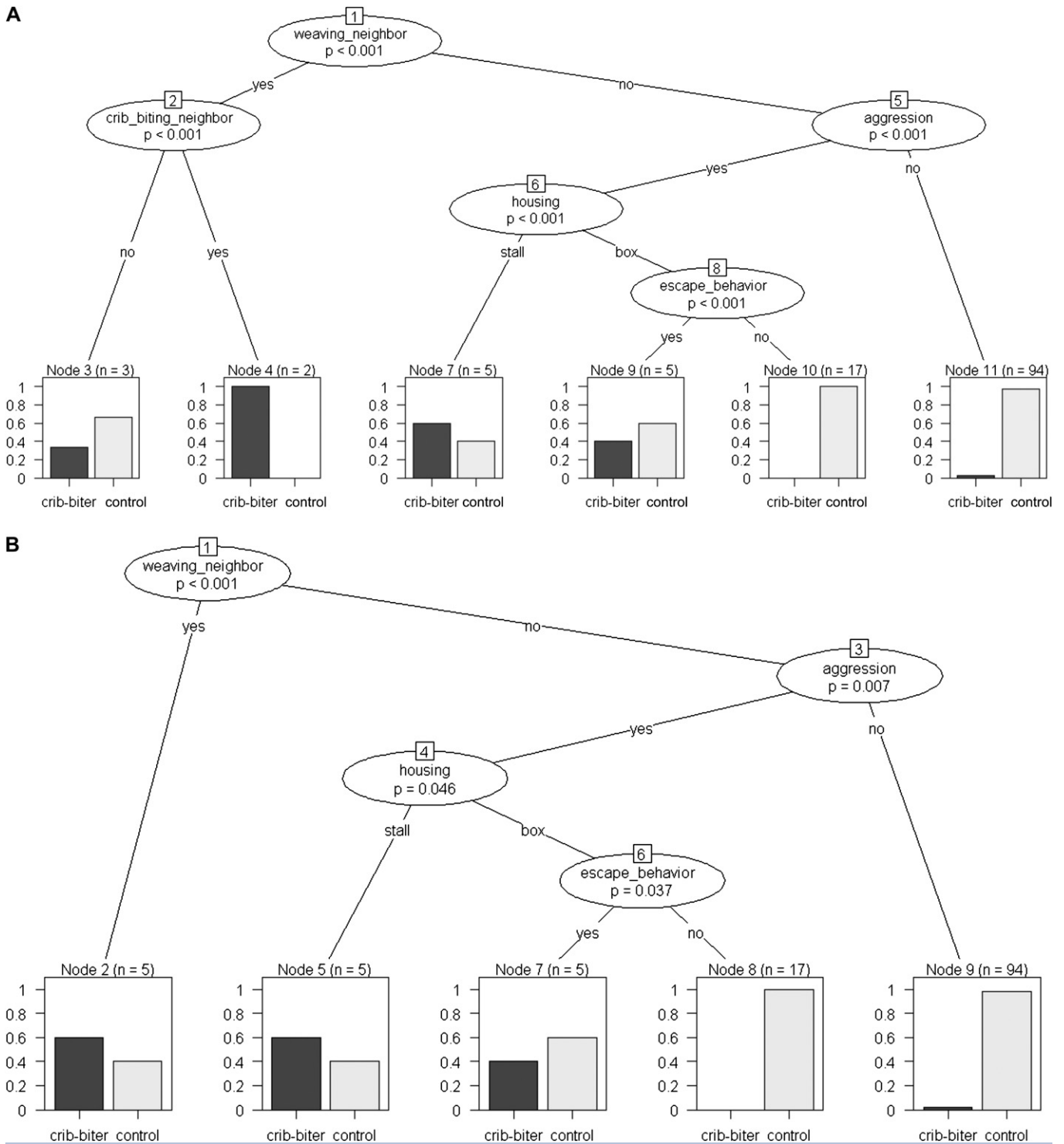


Figure 2. Conditional inference tree without (A) and with (B) correction. Risk factors selected by the algorithm are symbolized as ramification points of the conditional inference tree along with the uncorrected (A) or corrected (B) P -values. (A) Nodes 3, 4, 7, 9, 10, and 11; and (B) nodes 2, 5, 7, 8, and 9 are the terminal nodes. Bar plots visualize the probability of crib-biting behavior and the number of observations within each subgroups.

Classification Tree. Figure 1 shows the variables selected by the CART model before (Fig. 1A) and after (Fig. 1B) pruning. The first ramification point represents the first risk factor selected by the procedure. It tells us that the prevalence of crib-biting behavior is higher amongst horses that have a stereotypic (weaving) neighbor compared to those who do not have any. The left route represents altogether $2 + 1 + 0 = 3$ horses without and $0 + 0 + 2 = 2$ horses with crib-biting behavior, that is, 60% of those non-competition horses that had stereotypic (weaving) neighbor showed crib-biting behavior. In this route the next ramification point, that is, the next predictive factor found by the procedure was presence of a crib-biting neighbor. By contrast, horses that did not have a weaving neighbor were further divided by the model according to whether they show aggression behavior or not. Horses that did not have a weaving neighbor but were showing aggression toward other horses seemed to have higher prevalence of crib-biting behavior than those horses that were classified as not aggressive toward other horses by the owners. In other words, 19% ($3 + 1 + 1 + 0 = 5$ out of $5 + 1 + 4 + 17 = 27$) of those horses that were lacking of a stereotypic neighbor but had aggressive tendencies toward other horses showed crib-biting behavior. To improve homogeneity of the child nodes, two more risk factors were selected by the final pruned tree model (Fig. 1B). Out of those horses that showed aggression toward other horses, crib-biting behavior seemed to be more prevalent amongst horses housed in stalls (tethered with a rope in the stall, with restricted free movement), or housed in boxes but showing “escape behavior” (door opening or knot-untying). The remaining crib-biting horses were amongst horses that had no weaving neighbor and showed no aggression toward other horses. Classification tree method (after pruning) classified 97% of horses without stereotypic behavior and 60% of crib-biting horses correctly.

Conditional Inference Tree. Building a tree without any correction included more variables and nodes than the tree built using Bonferroni correction (Fig. 2A). Similarly, the pruned classification tree, the conditional inference tree with Bonferroni correction contained four risk factors and six nodes (Fig. 2B). The interpretation of this tree is similar to that of the CART model. Conditional inference tree with Bonferroni correction classified 97% of horses without stereotypic behavior and 60% of crib-biting horses correctly.

DISCUSSION

The prevalence of crib-biting found in this study was not different from those observed in other countries.¹⁻³ Analyses revealed similar risk factors in all three models. Contrary to the findings of Hothorn et al,⁹ we found no difference

regarding tree structure or predicting accuracy between classification tree and conditional inference tree methods. Berzal et al¹⁸ also reported that different splitting criteria had no significant effect on the accuracy of the classifier, and no single-splitting criterion proved to be universally better than the rest.

Risk factors revealed by logistic regression were less in quantity compared with the number of risk factors selected by CART and conditional inference tree methods, but all three methods found the two main risk factors reported previously by Nagy et al¹⁶: presence of a weaving neighbor and aggression toward horses. According to tree-based methods, it seems like that the influence of crib-biting neighbors on crib-biting behavior in horses may manifest itself only in special circumstances, which is also suggested by Albright et al.¹ Other risk factors selected by the tree-based methods are also in accordance with previous findings. Tethering as a method of managing horses is unsatisfactory from many points of view,¹⁹ and our results also suggest that if stable environment has a restrictive nature from a locomotor perspective (horses housed in stalls and tethered with a rope), horses are more likely to show crib-biting behavior.

Some risk factors reported by others, such as the effect of breed, were not identified in our study. This could be due to the small variability of breeds in our study.¹⁶ Most horses were Hungarian half-breeds, whereas the number of thoroughbreds, the breed most likely to be affected by stereotypies,¹ was as few as six. However, other studies also reported no difference between breeds with respect to stereotypic behavior.³ Some other factors often mentioned in association with crib-biting behavior, such as gastrointestinal discomfort⁴ or basal ganglia dysfunction,²⁰ were not examined in this study.

In comparison with the logistic regression, tree-based methods, due to their hierarchical nature, are able to demonstrate factors that are risk factors only under special conditions. However, the same factor can appear twice or more in the same tree (like crib-biting neighbor in Fig. 1A), and may have a slightly different role at each appearance, depending on the context, that is, the preceding factors located above it. This feature allows for detecting nonlinear relationships and interactions between the factors. Because of this flexibility, the tree method helps in better understanding of the problem under study. In our case, it also resulted in better prediction accuracy than the logistic regression. Specifically, prediction accuracy of crib-biting horses (sensitivity) was much better by the tree-based methods than by the logistic regression; however, specificity was only slightly lower. This is in accordance with previous findings.^{10,11,13}

Tree-based methods do not have strict applicability conditions like the logistic regression, work well with complex datasets, are less influenced by the multicollinearity of the

variables, and handle the missing values and low prevalence easily. Their output, the tree diagram, shows the probability of the occurrence of the events and vividly illustrates the structure of the risk factors and their complex interactions, which would be difficult or even impossible to model by logistic regression. Therefore, it makes the findings easier to interpret, even to those with less statistical background.^{7,12,13}

In summary, we can conclude that tree-based methods (either CART or conditional inference tree) are useful tools in finding risk factors, or even for data mining, alone or together with logistic regression method.

ACKNOWLEDGMENTS

The authors thank the support of Dr. G. Bodó, as well as K. Gavaldà and Z. Varga for assistance in the fieldwork.

REFERENCES

- Albright JD, Mohammed HO, Heleski CR, Wickens CL, Houpt KA. Crib-biting in US horses: breed predispositions and owner perceptions of aetiology. *Equine Vet J* 2009;41:455–458.
- Bachman I, Audigé L, Stauffacher M. Risk factors associated with behavioural disorders of crib-biting, weaving and box-walking in Swiss horses. *Equine Vet J* 2003;35:158–163.
- Parker M, Goodwin D, Redhead ES. Survey of breeders' management of horses in Europe, North America and Australia: comparison of factors associated with the development of abnormal behaviour. *Appl Anim Behav Sci* 2008;114:206–215.
- Waters AJ, Nicol CJ, French NP. Factors influencing the development of stereotypic and redirected behaviours in young horses: findings of a four year prospective epidemiological study. *Equine Vet J* 2002;34:572–579.
- Hosmer DW, Lemeshow S. *Applied logistic regression*, 2nd ed. New York, NY: Wiley; 2004.
- Kass G. An exploratory technique for investigating large quantities of categorical data. *Appl Stat* 1980;29:119–127.
- Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl* 2008;34:366–374.
- Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. New York, NY: Chapman & Hall; 1984.
- Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 2006;15:651–674.
- Worth AP, Cronin MT. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *J Mol Struct (Theochem)* 2003;622:97–111.
- Ho SH, Jee SH, Lee JE, Park JS. Analysis on risk factors for cervical cancer using induction technique. *Expert Syst Appl* 2004;27:97–105.
- Harper PR. A review and comparison of classification algorithms for medical decision making. *Health Policy* 2005;71:315–331.
- Thomas S, Leese M, Walsh E, McCrone P, Moran P, Burns T, et al. A comparison of statistical models in predicting violence in psychotic illness. *Compr Psychiatry* 2005;46:296–303.
- Low M, Joy MK, Makan T. Using regression trees to predict patterns of male provisioning in the stitchbird (hihi). *Anim Behav* 2006;71:1057–1068.
- Kubinyi E, Turcsán B, Miklósi Á. Dog and owner demographic characteristics and dog personality trait associations. *Behav Processes* 2009;81:392–401.
- Nagy K, Schrott A, Kabai P. Possible influence of neighbours on stereotypic behaviour in horses. *Appl Anim Behav Sci* 2008;111:321–328.
- R Development Core Team. R, A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: <http://www.r-project.org>; 2007.
- Berzal F, Cubero JC, Cuenca F, Martín-Bautista MJ. On the quest for easy-to-understand splitting rules. *Data and Knowledge Engineering* 2003;44:31–48.
- National Joint Equine Welfare Committee and the R.S.P.C.A. Code of practice for tethering horses and ponies. Available at: <http://www.rspca.org.uk/>.
- McBride S, Hemmings A. A neurologic perspective of equine stereotypy. *J Equine Vet Sci* 2009;29:10–16.