

Glossary of terms

provided by L. BODEN and T. PARKIN to accompany

Risk factors for Thoroughbred racehorse fatality in flat starts in Victoria, Australia (1989–2004)

L. A. BODEN*, G. A. ANDERSON, J. A. CHARLES, K. L. MORGAN†, J. M. MORTON‡, T. D. H. PARKIN§, A. F. CLARKE and R. F. SLOCOMBE

Department of Veterinary Science, The University of Melbourne, 250 Princes Highway, Werribee, Victoria 3030, Australia; †The Epidemiology Group, Department of Veterinary Clinical Science, University of Liverpool Veterinary Teaching Hospital, Leahurst, Neston, Cheshire CH64 7TE, UK; ‡School of Veterinary Science, The University of Queensland, Brisbane, Queensland 4072, Australia; §The Animal Health Trust, Lanwades Park, Kentford, Newmarket, Suffolk CB8 7UU, UK.

Association: Statistical dependence between 2 or more events, characteristics, or other variables. The presence of an association does not necessarily imply a causal relationship.

Bias (systematic error): Deviation of results or inferences from the truth, or processes leading to such deviation.

Biological plausibility: An association (or relationship between 2 factors) that is consistent with existing medical or veterinary knowledge.

Categorical data: Integer data with 2 or more exclusive categories that are counted rather than measured.

- Binary (Dichotomous) data:** Data with only 2 exclusive categories e.g. alive/dead, high/low.
- Polytomous data:** Data with more than 2 exclusive categories e.g. fast, good, dead, slow or heavy track surfaces.
- Nominal:** Data values consist of scores that have no inherent ordering e.g. breed, gender (female, male, neutered).
- Ordinal:** Data values consist of scores that are inherently ordered e.g. disease severity 0, 1+, 2+, 3+ or high/moderate/low.

Case-control studies: Comparison of exposures (potential explanatory variables) of individuals with disease (*cases*) with those of individuals without the disease (*controls*)

Causal diagrams: Enable the visualisation of the relationships between multiple factors and an outcome.

Clustering: Nonindependence of data either through common environment, spatial or geographic proximity, or through repeated measurements (Dohoo *et al.* 2003). For example, where the difference (variance) between horses within a cluster (e.g. training yard) is less than that between horses in different clusters or training yards.

Cohort studies: Comparison of 2 groups within a population (exposed and unexposed to potential risk factors) with respect to the development of the outcome under investigation.

Collinearity: A situation where there is a linear relationship between some or all of the independent (explanatory) variables in a regression model.

Confidence Interval (CI): Confidence intervals (CI) reflect the precision of the point estimate and indicate the range of values that a parameter is likely to take (Dohoo *et al.* 2003). If a 95% CI includes the null value (i.e. one for risk ratio, incidence rate ratio and odds ratio), it indicates that the point estimate is not statistically significant from the null value at a P value of 0.05 (Dohoo *et al.* 2003). Studies with larger sample sizes will tend to have narrower confidence intervals and the precision of point estimates will be greater.

Confounding variable, Confounder: A variable that can cause or prevent the outcome of interest, is not an intermediate variable, and is also associated with the factor under investigation. Unless it is possible to adjust for confounding variables, their effects cannot be distinguished from those of factor(s) being studied. Usually the effects of a potential confounder on the estimates for variables in a model are assessed by fitting each one at a time into the final model. If addition of the variable alters the point estimate for any of the variables in the final model by more than 20% (Dohoo *et al.* 2003), confounding is considered to be present, the confounder should be retained in the final model and adjusted point estimates reported for variables in the final model.

Continuous data: Data based on a continuous scale of measurement, such as age or weight, that is not restricted to integer values and that is measured rather than counted. Continuous data can be converted to discrete data by rounding and to categorical data by establishing cutoffs and classifying it into groups.

Correlation coefficient (r): The Pearson's correlation coefficient is the extent to which the association between 2 variables can be described by a straight line.

Cross-sectional studies: Involve the sampling of each study subject once at a specific point in time. They are most useful for

assessing permanent risk factors (for example, gender) (Dohoo *et al.* 2003) and for determining the prevalence of disease at a particular time point.

Deviance: A measure of the extent to which a model differs from the saturated model (a model in which there would be one parameter for every data point) (Everitt 2002).

Discrete data: Data that can be categorised into a classification. Discrete data is based on counts. Only a finite number of values are possible, and the values cannot be subdivided meaningfully. For example, the number of horses in a race can only be a whole number - it is not possible to have 12.3 horses in a race.

Diagnostics: Procedures for identifying departures from assumptions when fitting statistical models (e.g. residuals, Hosmer-Lemeshow goodness-of-fit test statistic, leverage, delta χ^2 and delta deviance).

Goodness-of-fit measures: Measures of agreement that provide an overall assessment of how well the model fits the observed data.

Interaction term: A term applied when 2 (or more) explanatory variables do not act independently on an outcome. Interaction is often referred to as effect modification where the size of effect of one explanatory variable on an outcome is dependent on the level of another explanatory variable.

Logistic regression: A form of regression to determine the relationship between one or more continuous or categorical explanatory variables and a binary outcome variable (live/dead, sick/well).

Multivariable: Refers to more than one explanatory variable. Multivariable regression refers to regression of more than 2 explanatory variables on an outcome. This can sometimes be mistaken with multivariate regression which describes models that have more than one outcome or response variable.

Odds ratio: A measure of the degree of association; for example, the odds of exposure among the cases compared with the odds of exposure among the controls. If the odds ratio (OR) = 1 there is no association if OR is significantly >1, there is a positive association with the outcome (a potential risk factor), if OR is significantly <1 there is a negative association with the outcome (a potentially protective factor).

Outcome variable (response variable, dependent variable): The variable on which the effects of explanatory (predictor) variables are being studied.

Power: Power is the likelihood that a study will detect a true difference of a given magnitude between groups if it actually exists (i.e. a true positive). Power is a function of study sample size, the biological variability in the population, the desired proportions of false positives (alpha) and false negatives (beta), and the type of statistical test used. Typical power levels are 0.80

and 0.90. The concept of power is extremely important because the lack of it (often due to small sample size) can lead to statistical insignificance even though true biological significance exists.

Predictor variable (explanatory, exposure, risk factor, independent variable): A variable that may predict or explain the outcome or response variable.

Regression analysis: Regression analysis models the relationship between one or more response variables (usually named Y), and the predictors (usually named X_1, \dots, X_p).

Receiver Operating Characteristic (ROC) curve: A plot of the sensitivity of a test against one minus specificity of a test. The closer the curve is to the upper left hand corner of the graph (area under the curve = 1.0, sensitivity and specificity 100%), the better predictive ability of the model. If the curve is close to the diagonal line (area under the curve = 0.5), the model has little predictive ability (Dohoo *et al.* 2003).

Residuals: The difference between the observed (true) value of an outcome variable and the value predicted by the model.

Risk factor: Individual characteristics or factors associated with an increased probability of developing a condition or disease.

Sample: A sample is a group of individuals that is a subset of a population and has been selected from the population in some fashion (e.g. random or convenience).

Sampling frame: The portion of the population from which the sample is selected.

Statistically significant: The conclusion that the results of a study are not likely to be due to chance alone because the P value derived from the statistical analysis is smaller than the critical alpha value (usually 0.05).

Stepwise regression: One of a series of methods for selection of a subset of explanatory variables when using regression analysis. Variables are entered into a model one at a time but after each addition of a new variable, those variables in the model are considered for removal by a backward elimination process.

Univariable (univariate): Data involving a single explanatory variable. The term is usually referred to in the context of univariable screening or analysis where the association between an explanatory variable and an outcome is assessed for each variable at a time. This step of an analysis does not consider the potential for confounding effects of other variables.

References

- Dohoo, I., Martin, W. and Stryhn, H. (2003) *Veterinary Epidemiologic Research*, AVC Inc., Charlottetown, Prince Edward Island.
- Everitt, B.S. (2002) *The Cambridge Dictionary of Statistics*, 2nd edn., Cambridge University Press, Cambridge.