Putting together the web of factors associated with complex diseases

D.U. Pfeiffer

Department of Veterinary Clinical Sciences, Massey University, Palmerston North, New Zealand

Veterinary epidemiology is the science of diseases in animal populations. It provides a methodology for investigating disease or production problems which are influenced by a large number of different factors or determinants. The term *web of causation* has been used to describe such systems. Under most circumstances disease is unlikely to be just dependent on the presence or absence of a single organism (*chain of causation*). It is more likely to be subject to a complex web of interacting factors involving the agent, host and environment (the *epidemiological triad*). The traditional clinical approach to veterinary disease and production problems is appropriate as long as the clinician is dealing with a single clearly identifiable organism causing a disease. But a lot of the problems a clinician is confronted with are more complex as they involve multiple factors and determinants of disease or production. Under such circumstances the task for the clinician is to assess the *presence* and *importance* of many *known* or *unknown* risk factors and their *interrelationships* in an *objective* way. This *multifactorial* view of disease or production effects is difficult to process in the human mind. Epidemiological tools can be used very effectively in such situations.

Study design

Observational studies are the epidemiologist's tool for investigating disease or production problems which involve a complex web of interacting, potentially causal factors. There are three main types of studies which could be used: *cross-sectional*, *case-control* and *cohort* studies (Kennedy 1990; Martin *et al* 1987). In a *cross-sectional* study a random sample of a population is taken at one point in time and examined for the presence of disease and associated risk factors. A *case-control* study is based on a sample of *cases* and *controls.* The *cases* include for example herds or animals with a disease and the *controls* should be selected from the same population as the *cases*, but in this example should not have the disease. Information about relevant risk factors is collected for each of the *cases* and *controls* In a *cohort* study a sample of herds or animals which are disease-free at the beginning of the study and which have been divided into cohorts according to a set of risk factors is followed over a given period of time. Disease occurrence is measured and compared between cohorts. In one way or another any such observational study uses a random sample of the *target population.* For the study to be successful it is imperative that it is planned meticulously. The investigator should be clear about his/her objectives, what the *target population* for his/her inferences is, what *sample size* is required to produce meaningful results, what information to collect, how much it is going to *cost*, how long it is going to take, how easy it will be to get cooperation from animal owners, how to store and finally how to analyse the data. As *data storage* and *analysis* require the use of computers, access to and competence in the use of these tools is essential.

Audigé *et al* (1993) can be used as an example of an epidemiological study in the deer health and production. Chesterton *et al* (1990) provide an example of the use of multivariate analysis techniques for disease problems in dairy cattle.

Data collection

During the field study, data will be collected from a sample of a population on the factors (*risk factors*) which are considered important for describing the disease or production problem (*outcome variables* of interest). The investigator has to make sure that the statistical sample satisfies the requirements of the type of study conducted. The data collected should be as complete as possible. Quite often researchers find out later that they had omitted obtaining data on a particularly relevant factor. The *data collection* phase is usually very costly and time consuming.

Data analysis

The objective of the *data analysis* phase is to extract meaningful information from the raw data. Typically, the first step in the analysis is to conduct a *descriptive analysis* which provides a descriptive overview of the system under study. This includes the calculation of statistical summaries as well as the tabular and graphical representation of the data. During this phase potential errors in the data can be identified and corrected. This is a very important step and usually provides the investigator with a useful quantitative description of the system such as for example the prevalence or incidence of specific disease.

Statistical analysis

So far no attempt has been made to look at *relationships* between potential risk factors (such as breed of an animal) and the outcome variable of interest (such as infection status of an animal). The objective in this case could be to find out if the *probability* of infection of an individual animal is dependent on its breed. In other words the question to be asked would be "Is the risk of infection for an individual animal any different if the animal belongs to breed A or breed B ?". If there is *dependence* between breed and infection status, a comparison of the two variables (infection status and breed) using the data collected during the study should show a difference between the proportion of diseased in animals of breed A and the proportion of diseased in animals of breed B which is unlikely to have occurred by *chance*. Statistical methods are used to quantify the probability that the observed difference is due to chance. In this example, a chi-square test could be used to test the relationship between the two variables for *statistical significance.* If the chi-square value was more than 3.84, then the associated p-value would be less than 0.05. This means that the observed difference between the two proportions would be expected to occur just due to chance variation less than 5 times out of 100 similar samples. It could be concluded that the two variables are *statistically significantly associated.*

In the following example it is assumed that the proportion of diseased animals is 0.30 in breed A and 0.50 in breed B animals. Using a 2-by-2 table a chi-square value of 8.33 with 1 degree of freedom and the associated p-value of 0.004 can be calculated (see Table 1). This p-value indicates that the observed difference between the two proportions would be expected to occur due to chance variation alone less than 4 times in 1000 similar samples. The result of this statistical analysis therefore allows the conclusion that the risk of infection in this population is not independent from breed and the two variables are statistically significantly associated. Hence, animals of breed A are less likely to become infected than animals of breed B.

Table 1 Comparison of risk of infection and breed

Breed	Infection status		
	positive	negative	Prevalence (proportion)
A	30	70	0.30
B	50	50	0.50

Univariate analysis

The first step in the statistical analysis is called the *univariate analysis* where each of the potential risk factors is tested for the *statistical significance* of its *association* with the outcome variable of interest. After this step in the analysis the researcher will know which factors are likely to be related to the outcome variable. It is not known which factors are most important, which factors are interacting and which are confounding factors. Hence, it is not yet possible to describe the *web of causation.*

An example for this approach can be found in the paper by Audigé *et al* (1994) elsewhere in this proceedings describing the risk factors for growth and reproduction in New Zealand deer herds.

Multivariate analysis

The final analysis step consists of the *multivariate analysis* (for an example see Audigé *et al* elsewhere in this proceedings). The objective of this analysis is to study the interrelationships between multiple risk factors with respect to their effect on the outcome variable. This should provide the investigator with a quantitative description of the *web of causation.* Knowledge about the importance of each risk factor and the relationships between factors may allow the investigator to develop a strategic approach towards controlling the disease or production problem.

The relationships between risk factors can be subject to *interaction* and *confounding* A factor may have a *direct* and/or an *indirect* effect on the outcome variable. These issues have to be investigated during the multivariate analysis. An *interaction* between two variables implies that the effect of each of the two risk factors on the outcome variable depends on the level of the other risk factor. An example would be a situation where the risk of respiratory disease varies depending on whether a viral or bacterial agent are present each by themselves or together. The risk may be much higher if both are present at the same time. *Confounding* is different in that it represents a situation where a potential risk factor appears to be related with the outcome variable, but in fact represents the effect of another risk factor which in turn it is associated with. An example from a study conducted in New Zealand suggested that the risk of leptospirosis infection in milkers increased if the milker was wearing an apron during milking. It turned out that milkers using an apron were also working with larger herds. Hence, wearing an apron was confounded with herd size, the latter being the actual risk factor.

Knowledge about indirect as well as direct effects is important as it may provide important clues in preventing a particular problem. For example the presence of a particular organism may be the direct cause of a particular disease. But the organism may only occur under certain environmental circumstances. Hence, the environmental conditions would have an indirect effect on the disease. Yet, the most appropriate way of controlling the disease could be by changing the environment if possible.

There is a wide range of statistical methodology available to conduct *multivariate analyses*. For the description of a *causal web* the method of *path analysis* seems to be most appropriate. This technique has the advantage that it combines knowledge and hypotheses about the underlying biological system with the power of statistical methods. The investigator first has to design a *null hypothesis* path diagram which represents the interrelationships between factors according to his/her biological understanding of the system studied. The diagram should be based mainly on the factors which came out as statistically significant in the *univariate analysis*. If there are any other factors which are considered biologically significant, they can be included in the appropriate place within the *null hypothesis path model*. The result of this process should be a diagram based on a number of factors linked through arrows which are thought to be *directly* or *indirectly* related to the outcome variable. The next step of the path analysis involves testing each of the hypothesized relationships in the *path model* for statistical significance. A statistical method called *multiple regression* (linear or logistic - depending on the type of dependent variable) is used to perform this analysis. Basically each variable (*dependent* variable) in the path model with arrows leading to it is regressed on the respective variables or risk factors (*independent* variables). Each arrow which does not turn out to be statistically significantly associated with the dependent variable is dropped from the path model. The importance of the individual factors can be quantified on the basis of their regression coefficients. *Confounding* effects can be controlled for in these analyses by including them in the respective regression models. *Interaction* effects between risk factors have to be determined by including them in the regression analyses.

The result of the *path analysis* should be a graphical representation of the *causal web* underlying the system studied.

What does it all mean ???

The final and probably most important step of the analysis will be the interpretation of the results. If a technique such as *path analysis* was used the investigator can use the graphical representation of the system to compare and communicate the required action in order to solve the disease or production problem. The researcher will be able to base any action on a more objective basis rather than "just" clinical intuition. Using the *path model* the factors in the *causal web* which are cost-effective to control can be identified. The model may generate new hypotheses which would have to be tested during a follow-up study.

But it should be kept in mind that this model still is only a representation of the data which was collected. How well it relates to the actual situation in the target population, depends on the representativeness of the sample and the quality of the data as it was collected and recorded. And it is likely that it is possible to construct a number of different models of the same system.

It has to be remembered that the *statistical analysis* of observational studies produces information about *statistical associations* between variables. It cannot prove *cause-effect relationships*. This would require that *random* and *systematic error* could be completely controlled, which to some extent can be done during *clinical* and *experimental trials*.

Analysis tools

With the widespread use of personal computers and the development of appropriate software the tools have become available which facilitate storage and analysis of the data which has been collected during an epidemiological field study. The World Health Organization in cooperation with the U.S. Centers for Disease Control (distributed by USD, Inc., 2075-A West Park Place, Stone Mountain, GA 30067, U.S.A.; FAX: 404 469 0681) has developed the software package *Epi Info* (now available as version 6.0) which includes a database management system, a comprehensive set of statistical analysis tools and even a simple word processor for the design of questionnaires. The package is in the public domain and can be freely distributed.

References

Audigé L, Wilson P.R., Morris R.S. : Deer Mortality Profile (1994). See pages 251-256 of these proceedings.

Audigé L, Wilson P.R., Morris R.S., Pfeiffer D.U. : Risk Factors for Weaner Deer Bodyweight (1994). See pages 318-326 of these proceedings.

Chesterton R.N., Pfeiffer D.U., Morris R.S. and Tanner C.M. : 1990 : Environmental and behavioural factors affecting the prevalence of foot lameness in New Zealand dairy herds - a case-control study. *N.Z. Vet.J.* 37, 135-142.

Kennedy D. (technical editor) 1990: *Epidemiological skills in animal health.* Proceedings 143, Post Graduate Committee in Veterinary Science, University of Sydney, Sydney, Australia, 409pp.

Martin S.W., Meek A.H. and Willeberg P. 1987 : *Veterinary Epidemiology - Principles and Methods.* Iowa State University Press, Ames, Iowa, U.S.A., 343pp.