Epidémiol. santé anim., 1997, 31-32

A STRUCTURED APPROACH FOR ANALYSING SURVEY DATA AND MAKING USEFUL CAUSAL INFERENCES

Martin W.[1]

Beaucoup d'enquêtes épidémiologiques de terrain cherchent à collecter des données relatives à un grand nombre de variables, souvent plus de variables qu'il n'y a de cas. Bien que des approches statistiques variées conduisent à l'élaboration d'un modèle, les résultats de celui-ci risquent d'être assez peu fiables en terme d'identification des causes réelles de la situation observée. Une approche souvent utilisée, le criblage univariable d'une liste de variables, est décevante. On recommande plutôt aux chercheurs d'être beaucoup plus sélectifs dans leur choix de quelques variables significatives bien définies et d'identifier les paramètres révélateurs majeurs ou les effets confondants. Ils doivent aussi consigner de manière explicite une structure causale et les relations causales attendues pour chaque variable significative. Par conséquent, l'analyse devrait refléter la structure causale, quelle que soit sa puissance, et une attention soutenue devrait être portée à la définition des variables pour éviter des problèmes de multicolinéarité. Une meilleure définition et une meilleure quantification des variables, accompagnées d'une analyse structurée, sont sans doute plus à même d'améliorer nos modèles multivariables que la collecte de données sur encore plus de variables.

Epidemiologists are, in large part, pragmatists. Hence when we conduct a research project our main objective is to identify factors, through a process of detecting statistically significant associations, which are causally related to the outcome of interest (Charlton, 1996) and which can be manipulated, at an acceptable benefit-cost ratio, to prevent or reduce the unfavourable outcome(s). In approaching this task there are a number of issues relating to study design that must be considered. In this paper I will focus on the issue of choosing the number and scope of variables for which information will be sought and propose ways of analysing these variables that will increase the likelihood of achieving our main objective.

Experimental scientists usually approach research problems with the intent of investigating the effects of one or two variables of interest and rely on restricted sampling and randomization to prevent confounding of their effects. While valid within the context of the experimental design, the concern is that the results may not apply to the "real world". Epidemiologists, perhaps with the implicit belief that disease has multiple causes (Charlton, 1996), often study one or two categories of variables (eg. management factors, housing factors, feeding/ration factors) as their major interest and include other categories of variables as potential confounders (eg. owner characteristics, animal attributes, etc.). Since each category of variable often contains a number of individual variables, many epidemiologic studies begin with the intent of measuring tens or even hundreds of independent variables (Dohoo et al, 1997). Despite its widespread usage, the underlying detrimental effects of this approach in terms limiting our ability to develop valid models and achieve our main objective cannot be ignored. Five issues relating to analysing a large number of independent variables were outlined in a recent paper which focused on problems arising from multicollinearity (Dohoo et al, 1997). In this paper, more emphasis will be placed on the actual number of variables relative to the number of cases although a number of the other issues will be mentioned as a basis for developing a structured approach to this problem.

PROBLEMS FROM A RELATIVE EXCESS OF VARIABLES

The major problem is how to ensure that real causal associations are identified, as statistically significant, while excluding as many false associations (or idiosyncratic variables) as possible when there is a large number of parameters to be estimated (Thomas et al,1985; Concato et al, 1993; Robins and Greenland, 1986). Typically, some form of multivariable modelling will be used to help achieve this goal, and while necessary to our success, the researcher must bring some prior knowledge about the causal role of variables to the analytic process if one hopes to develop valid (useful) models (Robins and Greenland, 1986). The larger the number of variables the more essential it becomes that our knowledge of causal roles of variables be used to guide our modelling activities. The situation is more complex than just considering the initial number of variables because some of the p' original variables may be categorical and need to be subdivided into a greater number of dummy variables; continuous variables may need to be entered as quadratic or higher transformations, and interaction terms need to be considered, all of which further increases the number of variables to be tested. Thus, the total number of parameters (p") that will be evaluated in the analytic process may be considerably larger than the original number of variables (p') would imply. Simulations suggest that the number of parameters estimated (ie. number of variables, including transformations, etc. (p")) should be less than n/10, where n is the number of cases (with a continuous outcome) or the number of cases in the less frequent outcome category (with a binary outcome). Thus, if there are 100 cases, with at least as many controls, in a logistic model, p" should be no greater than 10. Using too large a number of possible predictors usually results in over-fitting which is reflected in a flattening of

[1] Department of Population Medicine, Ontario Veterinary College, Guelph, Ontario CANADA N1G 2W1

the $Y\hat{Y}$ plot away from the 45° line, a pattern called shrinkage. The shrinkage coefficient ( ) is a multiplier of the XB matrices which will make the model as good for predicting in future data sets as it is in the primary or developmental data set. Roughly is (Model ^{2} - p") / Model ^{2} where model ^{2} is the likelihood ratio ^{2} statistic resulting when all p" parameters are fitted in the model (versus the model with intercept only). In linear regression the shrinkage coefficient (for the full model) is (Adjusted R^2) / R^2 (Harrell et al, 1996). For example, if = 0.8, then about 20% of the model fit is due to noise variables (regardless of the number of predictor variables included in the final model). Given that the maximum desired number of variables is p_{max} n/10, unless the Model X^2 p_{max} + 9 then variable reduction will not likely lead to a useful model. Unfortunately since often p">n, it is not possible to estimate directly, but this fact should make us aware of how tenuous our models may be and in this instance using the n/10 rule is likely the best approach.

Most epidemiologists recognize the need for reducing the number of variables, but most appear to do this mechanically after data collection as opposed to the preferred route of reducing the number at the planning stage. Indeed, using a univariable (one predictor variable) association approach is one of the most common ways of reducing the number of variables to be included in a subsequent multivariable model. However, this "screening" approach has distinct disadvantages because it ignores the effects of confounders (Sun et al, 1996), specifically the class of confounders called distorters (Susser, 1973). The most straightforward example of where distortion will occur is when the confounder has an association of opposite sign/direction, with the variable of interest, to the sign of their association with the outcome. The same effect occurs when the confounder and the variable of interest have an association whose sign differs from the sign of the association of either the variable of interest or the confounder with the outcome. Depending on the situation, one, or the other, or both variables may have a nonsignificant univariable association with the outcome. Thus, potentially important variables can be incorrectly omitted from the multivariable modelling process. Most choose to use a relaxed (type 1 error) level, say =0.15 or 0.25, which helps alleviate some of the problems with univariable screening it does not remove them. The univariable approach cannot differentiate between authentic and noise variables and a study design that focuses on a reduced set of variables that are known or strongly suspected of being risk factors for the outcome is preferable.

If post collection variable screening is needed, a better approach is to use bivariable or trivariable screening with the one or two most important confounders included to create strata in a Mantel-Haenszel type approach. Again, this returns us to the necessity of using a priori beliefs about causation to guide our analyses. Alternatively, if there are not too many variables for the number of cases then one could fit the full model and then use backward elimination, to achieve one "best" model, although some workers indicate that it is preferable to leave the full model intact and not use variable reduction. Using an all possible subsets approach to view a number of "good" alternative models may also be preferable to the single best model approach and often indicates why the choice of the best model is open to a great deal of uncertainty (Robins and Greenland, 1986; Harrell et al, 1996).

It would appear that most workers do not have a predetermined number of variables in mind when the data screening process is used, but if the number of variables is much greater than n/10, then one of the formal variable reduction techniques (discussed by Dohoo et al, 1997) should be used to reduce the number of confounder variables entered into multivariable models. This is much preferable to just omitting these variables. In general, one would exclude the key variable(s) of interest from these factors/components, unless the intent is to treat the subsets of factors (say ration) as a group and not individually. The basic principle is that variable reduction should only involve the X variables, not the Y variable as is done whenever a preliminary screening process is used.

Others prefer not to screen and choose to use a forward or stepwise forward selection process to select "significant" variables (with or without a formal penalty technique to counteract the multiple testing). The Bonferroni correction is essentially /p", although this may be viewed as too stringent an approach to variable selection.

Originally, the justification for investigating a large number of factors was mainly one of ignorance. Since we did not know what factors might be important, we want to look at all "sensible" possibilities. In fact it appears that we believe that we have no choice except to continue this approach for the foreseeable future (Dohoo et al, 1997). However, given that a number of these data dredging studies are completed, it is less defensible to continue this approach and it behoves all of us to design studies based on much more well defined, and fewer, variables. For a thorough discussion of multivariable modelling the reader should consult Harrel et all, 1996. One needs to consider the potential benefits of accurately measuring a few well defined variables, including the major confounders, and developing our models based on these, versus our traditional approach of having too many variables for the number of cases and hoping that a statistical selection process can provide us with a model which reflects reality.

SUGGESTED SOLUTIONS

Clearly, every study should begin with a thorough review of the literature and a formulation of specific questions to be addressed. Often, as a result of trying to measure numerous variables, our level of definition of these variables is quite general. This inevitably creates problems in attempting to interpret the resulting model(s) and thus we should refine our definitions to the greatest extent practical. Further, it is rare that we posit guesses about the sensitivity and specificity of our measuring instruments. While not the focus of this paper, these issues beg resolution (Gordis, 1979; Martin, 1996). One of the important nonquantitative ways of reducing the number of variables we study is to think about the causal role/function of each variable or set of variables. In fact it is useful to explicitly define the variable, and then draw or record the nature of the "expected association" as part of the planning stage (Martin, 1996). These conceptual causal structures can influence the subsequent choice and definition of variables that data will be collected on, they should influence the analytic approach as well as the interpretation of model results (Robins and Greenland, 1986), and in the extreme a formal causal structure will dictate the specific modelling process to be used (Buncher et al, 1991). A major advantage of using a formal structural model approach to analysis is that if forces us to consider both temporal sequence of variables and a priori plausibility arguments about causal roles.

If, at the planning stage, the number of variables that should be investigated remains large, then a process, such as the Delphi technique, using an expert panel to prioritize the list might be useful (Dewey et al, 1995). Certainly, every researcher must differentiate between variables of interest and potential covariates or confounders. In most studies there is no need to include variables as potential confounders unless they are strongly suspected of being causal factors for the outcome of interest. Following this rule alone will greatly reduce the number of variables we collect data on. As an informal starting point, and without considering a specific causal structure, for each proposed variable we should ask questions about how that factor (eg. the number of veterinary visits, owner experience, animal management practices, using hired help, etc) might impact (ie. cause) the outcome? If the value of the variable might be a result of, instead of a cause of, the outcome perhaps redefine or omit that variable. If the variable is unlikely to be a direct cause of the outcome consider omitting it (bear in mind that the "stepping packages" only identify directly causal variables anyway), or be prepared to undertake a more complex analysis of the data with a causal structure in mind. Given that many analyses are embarked on with no structures in mind, many intervening variables will be allowed to enter the models and this prevents the detection of important but more, causally, distant variables (Susser, 1973). Finally, we need to be careful in how we define and enter variables into our models, since in many circumstances the models are not sensible on face value, even before considering the biological details of the outcome (Dorfman and Kimball, 1985). For example, the model used to demonstrate the control of multicollinearity by Lafi and Kaneene, 1992, is arguably nonsensible whether principal component regression is used or not. In this instance variables which were exact combinations of other variables in the model were entered, much as happens with interaction terms. Although the latter approach to modelling interaction is necessary, the construction of highly correlated variables is to be discouraged whenever possible. After analyses are completed we can use the previously recorded a priori guess about the effects of variables and we might also be able to reject variables with (but still report) statistically significant results (Robins and Greenland, 1986). Otherwise we are often forced to explain nonsensible, but "statistically significant" associations. While including "significant" variables in the model can improve the fit and the apparent utility of the current model, we should not accept associations that we have reasonable biological grounds to reject, just for the sake of the improved fit. Despite the large number of variables we investigate our models often perform only moderately well at best (low R^2, low sensitivity/specificity),however better definition and measurement of variables is more likely to improve this than collecting data on yet more variables.

BIBLIOGRAPHY

Buncher C.R., Succop P.A., Dietrich K.N., 1991. Structural equation modeling in environmental risk assessment. Environmental Health Perspectives 90, 209-213.

Charlton B.G., 1996. Attribution of causation in epidemiology: chain or mosaic? Journal of Clinical Epidemiology 49, 105-107.

Concato J., Feinstein A.R., Holford T.R., 1993. The risk of determining risk with multivariable models. Annals of Internal Medicine 118: 201-210.

Dewey C.E., Martin S.W., Friendship R.M., Kennedy B.W., Wison M.R. 1995. Associations between litter size and specific herd level management factors in Ontario Swine. Preventive Veterinary Medicine 22, 89-102.

Dohoo I.R., Ducrot C., Fourichon C., Donald A., Hurnik D., 1997. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Preventive Veterinary Medicine 29, 221-240.

Dorfman A, Kimball AW. 1985. Regression modelling of consumption or exposure variables classified by type. American Journal of Epidemiology 122, 1096-1107.

Gordis L. - Assuring the quality of questionnaire data in epidemiologic research. American Journal of Epidemiology 109, 21-24.

Harrell F.E. Jr., Lee K.L., Mark D.B., 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15, 361-387.

Lafi S.Q., Kaneene J.B., 1992. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. American Journal of Epidemiology 13, 261-276.

Martin S.W., 1996. If Multivariable Modelling is the Answer, What is the Question? Dutch Society for Veterinary Epidemiology and Economics, Wageningen, (9), 1-6.

Robins J.M., Greenland S., 1986. The role of model selection in causal inference from nonexperimental data. American Journal of Epidemiology 123, 392-402.

Selvin H.C., Stuart A., 1966. Data-dredging procedures in survey analysis. The American Statistician June, 20-23.

Sun, G-W. Shook T.L., Kay G.L., 1996. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology 49, 907-916.

Susser M. 1973. Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology. Oxford University Press. Toronto. 181pp.

Thomas D.C., Siemiatycki J., Dewar R., Robins J., Goldberg M., Armstrong B.G., 1985. The problem of multiple interference in studies designed to generate hypotheses. American Journal of Epidemiology 122, 1080-1095.