A MODIFIED CASE-CONTROL APPROACH TO IDENTIFYING RELATIONSHIPS BETWEEN DISEASES

Ian R. Dohoo[1]
S. Wayne Martin[2]

In order to evaluate the effects of a disease on productivity it is essential to understand the relationships between the disease of interest and other diseases of that species. Consequently, it is necessary to investigate the role of various diseases as determinants of other diseases.

In addition, in order to establish causation in observational studies, the time sequence of two events must be known with some certainty. This requirement, although logical, is not easily fulfilled, particularly in a study involving many diseases. In the past, studies have either identified associations between diseases without regard to the time order of those diseases or have evaluated relationships based on several possible time orderings without attempting to identify one order as correct. Relationships between two diseases can be of the form:

	I.	A ---> B	(A causes B)
	II.	B ---> A	(B causes A)
or	III.	C ---> A, C ---> B	(C is a common cause of A and B. There may or may not be any predisposition for A to precede B or vice versa (Susser, 1973))

where A refers to the putative causal disease and B refers to the putative outcome disease. Relationships established solely on the basis of an association between two diseases within a lactation may incorporate all three possible

1. Animal Pathology Division, Food Production and Inspection Branch, Agriculture Canada, Ottawa Canada.
2. Dept. of Veterinary Microbiology and Immunology, Ontario Veterinary College, Guelph, Ontario N1G 2W1 Canada.

forms, but only relationships of the first form satisfy the criteria for A to be considered a cause of B.

With this in mind, a technique which took into account the time to first diagnosis of two diseases was developed and used to evaluate relationships amongst pairs of diseases. The time ordering within a lactation of the major diseases of dairy cattle was postulated based on the observed median times to first diagnosis after parturition and apriori knowledge of disease relationships. Each disease in turn was considered an outcome disease and all diseases prior to it in the time ordering were considered as possible causal diseases.

The data base was derived from a study undertaken to evaluate relationships amongst diseases and production parameters in dairy cattle. The study involved 32 dairy herds within a sixty mile radius of Guelph, Ontario and spanned a period of two and one half years.

METHODOLOGY

A total of 1834 lactations were used in this study and a computer program was written in APL (A Programming Language) to facilitate organizing and analyzing these data. For each farm, all cows which developed the outcome disease of interest were identified and referred to as cases. From the remaining cows in that herd, the largest possible group of controls which contained an exact multiple of the number of cases was randomly selected. The controls were then subdivided into equal sized groups with one group being assigned to each case. The number of cases which had the putative causal disease prior to the day on which the outcome disease was diagnosed was recorded. Similarly, the number of controls having the causal disease prior to the time postpartum at which the outcome disease was diagnosed in their associated case was recorded. This ensured that only occurrences of the putative causal disease occurring prior to the outcome disease contributed to the association. In addition, it ensured that the case and its assigned group of controls were observed for the same period of time.

The program generated a fourfold table for each of the 32 farms. All tables with marginal zeros were excluded from subsequent analyses since they contributed nothing to the summary odds ratio estimate (Breslow and Day, 1980). If less than eight tables remained no further analyses were

performed and it was concluded that no relationship could be found between the two diseases under investigation. Consequently, at least eight farms had to have had at least one occurrence of each of the two diseases before a relationship between the two could be found.

If at least eight tables remained, the data from these tables were pooled to obtain a summary odds ratio using the Mantel and Haenszel technique (Mantel and Haenszel, 1959). This procedure combines the odds ratios from the individual tables into an estimated summary odds ratio by weighting the individual tables according to their inverse variances. The technique is not affected by zero cell entries and gives consistent estimates of the summary odds ratio even in the presence of small strata (individual tables) (Breslow and Day, 1980). The significance of the summary odds ratio was tested using a chi-square statistic at $p=0.1$. An algorithm for performing the required calculations for the case-control procedure is given in Appendix 1.

The analysis contained 17 disease conditions so there were 136 possible relationships within the intial postulated time ordering. This procedure was applied to each of those relationships. For comparative purposes, associations were evaluated by simply crosstabulating the occurrence of each of the two diseases within a lactation, ignoring the possible confounding effects of herd and without regard to the time of onset of the two diseases.

RESULTS AND DISCUSSION

The crosstabulation procedure identified 32 relationships as being significant at $p=0.1$. The modified case-control procedure identified 17 relationships as significant, 16 of which had been identified by the crosstabulation procedure. The one additional association identified by the case-control procedure (dystocia --> cystic ovaries) presumably resulted from different methods of calculation of the chi-square statistic since the time ordering of the relationship is unambiguous. In the crosstabulation procedure the chi-square statistic fell just short of significance while in the modified case-control procedure it just obtained it.

There are several reasons why the modified case-control procedure is more conservative than the crosstabulation procedure. The first relates to the time of occurrence of the diseases, since the crosstabulation procedure identi-

fies relationships of all three types presented above whereas the modified case-control approach only identifies relationships of type I or type III (provided there was a clear predisposition for A to occur before B). Relationships in the form of type III would subsequently be identified (if variable C represented a third disease) by identifying the relationships C --> A and C --> B. In this study associations between ketosis and five other diseases (teat injury, digestive disorders, metritis diagnosed at 21 to 60 days postpartum, mastitis and cystic ovaries) were identified as significant by the crosstabulation procedure but not by the case-control approach. All of these five diseases followed ketosis in the postulated time ordering so the results of the case-control analysis suggest that ketosis was not a determinant of any of those diseases. The crosstabulation procedure may have identified relationships in which ketosis was secondary to the other clinical diseases.

The second reason relates to the intensity of observation of case and control cows. Failure to provide observation of equal intensity in the two study groups may result in the identification of spurious associations (MacMahon and Pugh, 1970). In this study, it is likely that cows experiencing the outcome disease (ie: the cases) were likely to receive a thorough veterinary examination for that condition. Consequently there was a greater probability of the causal disease being diagnosed in those cows than there was in control cows and this may have resulted in spurious associations being found in the crosstabulation analysis. The modified case-control procedure circumvented this problem by only recording causal diseases if they occurred at least one day prior to the diagnosis of the outcome disease. This ensured that the intensity of observation for the causal disease was equal in the two groups of cows. As mentioned previously, there were five relationships involving ketosis identified as being significant by the crosstabulation procedure but not by the case-control procedure. As discussed, this may have been due, in part, to the time ordering in which those conditions occurred. However, it is also likely that the probability of ketosis being diagnosed in cows being examined and treated for the other conditions was higher than in the control cows.

The third reason relates to the sample size and its distribution within the fourfold table and the effect that this has on the calculation of the chi-square statistic in each procedure. The cell of the fourfold table with the smallest expected number of individuals was the cell in

which the cows had experienced both the putative causal disease and the outcome disease. For a fixed sample size, incorporation of more animals into this cell would increase the magnitude of the chi-square statistic. The crosstabulation procedure would likely have incorporated more individuals into that cell. In addition, the crosstabulation procedure had a slightly larger sample size because during the selection of controls in the case-control procedure, a few animals from each herd were excluded. Eight of the sixteen relationships which were identified as significant by the crosstabulation procedure but not by the case-control procedure were based on the two diseases occurring together in eight or less cows.

The primary reason for developing the modified case-control procedure was to evaluate relationships amongst diseases to assist construction of a model of diseases suitable for path analysis. Observations about disease occurrence were stratified according to herd of origin but not by age at calving. Consequently, age may have been a confounding variable in the evaluation of these relationships. Since these relationships were being used as a basis for a path model in which age was included as an exogenous variable, the effects of age would be controlled in that analysis. However, if a larger sample was available, it would be feasible to stratify on the basis of age as well as herd.

Only relationships which conformed to the postulated time ordering of diseases were examined in this study. It would be possible to apply this technique to all possible pairs of diseases with each disease being investigated as both a cause and effect of the other disease. The results of this approach could subsequently be used to postulate a time ordering of the various diseases. However, it is possible that relationships of the form A-->B and B-->A may be identified as significant and this would require modifications in the subsequent path analysis.

It is important to realize that this procedure was applied to the time of first diagnosis for the various disease conditions, which may vary somewhat from the time of onset of the diseases. As it was used, the technique was evaluating the role that one disease exhibiting detectable signs played in the development of detectable signs of another disease. As methods of detecting disease closer to their time of onset are developed and adopted, the procedure will more truly evaluate the role of one disease in the initiation of other diseases. Finally, with the increase in computerized data banks of animal diseases in domestic

animals, most of which will store disease occurrence by date, this procedure will be useful in the identification of relationships amongst those diseases.

REFERENCES

1. Susser, M., 1973. Causal Thinking in the Health Sciences. Concepts and Principles of Epidemiology. Oxford University Press, Toronto, 181 pp.

2. Breslow, N.E. and Day, N.E., 1980. Statistical Methods in Cancer Research. Volume 1. The Analysis of Case-Control Studies. International Agency for Research on Cancer, scientific publication 32, Lyons, 338 pp.

3. Mantel, N. and Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease, J. Nat. Cancer Inst., 22:719-748.

4. MacMahon, B. and Pugh, T.F., 1970. Epidemiology Principles and Methods. Little Brown and Co., Boston, 376 pp.

APPENDIX 1

Algorithm for Modified Case-Control Procedure

1. Obtain data for first farm. Data must consist of number of days postpartum to both the outcome disease and the causal disease and it must use a unique value to indicate if the disease was not diagnosed during the lactation.
2. Determine which cows had the outcome disease (ie: the cases)
3. Determine which cows did not have the outcome disease (ie: the controls)
4. Randomly select controls so that the number of controls is the largest possible exact multiple of the number of cases.
5. Divide controls into equal sized groups with one group being assigned to each case.
6. Count the number of cases which had the causal disease at least one day prior to the outcome disease (Cell A of the fourfold table)
7. Count the number of cases which did not have the causal disease at least one day prior to the outcome disease (Cell B)
8. In the first group of controls, count the number of cows which had the causal disease at least one day before the time postpartum at which its associated case had the outcome disease diagnosed.
9. Repeat step 8 for each of the remaining groups of controls and sum the results from all groups (Cell C)
10. Subtract Cell C from the total number of controls to get Cell D.
11. Repeat steps 1-10 to generate a fourfold table for each farm on the study.
12. Use the Mantel and Haenszel procedure to obtain an estimate of the summary odds ratio.