MULTIVARIATE MODELS IN DISEASE CONTROL PLANNING

Tsegaye Habtemariam* and Roger Ruppanner*

The epidemiology of complex infectious processes could be elucidated best when the multifactorial components of the infections are interrelated qualitatively and quantitatively. The use of multivariate analytic approaches to establish the quantitative relationships amongst a set of multicausal variables is appropriate for such problem solving tasks. Based on multivariate analytic models, realistic disease control approaches could be developed and rational recommendations made. Such an example was recently applied to the case of African trypanosomiasis in Ethiopia (Habtemariam 1979).

The trypanosomiasis-tsetse problem is complex and multicausal involving determinants of diverse nature. These include people, livestock and game populations, and geographic and climatic variables (Ford 1971; Glover 1967; and Wilson et al, 1975). The prevalence of trypanosomiasis in cattle and of its vector (Glossina spp.) is influenced by these diverse determinants either directly or indirectly.

To answer the question: what analytic methodology can be used to analyze and quantify the complex relationships among the determinants and prevalence of trypanosomiasis, quantifiable information must be available. This information, when examined with the appropriate analytical method, would enhance the understanding of the epidemiology of the disease and also provide a means for formulating a trypanosomiasis control program.

The objective of the present study was to derive mathematical functions that could be utilized: a) to explain the underlying biological process in the epidemiology of the trypanosomiasis-tsetse complex quantitatively, b) to obtain a subset of predictor variables that will provide the best linear prediction equation for the prevalence of trypanosomiasis in a given area, and c) to classify a given area as a high or a low trypanosome risk area. Multiple regression and discriminant analysis, two multivariate techniques best

*Department of Large Animal Medicine, School of Veterinary Medicine, Tuskegee Institute, Alabama 36088 (Habtemariam); Department of Epidemiology and Preventive Medicine, University of California 95616 (Ruppanner)

suited for studying such problems were selected; detailed presentations of the mathematical algorithms involved and their applications to practical problems are available elsewhere (Morrison 1976; Bolch and Huang 1974; and Neter and Wasserman 1974).

Materials and Methods

An epidemiologically relevant causal diagram was developed to facilitate the rational decomposition of sequences of relationships and/or interactions amongst variables that influence trypanosomiasis (Fig. 1). With the aid of this diagram, 28 predictor variables were identified and grouped into determinants which influence the dependent variable, the prevalence (%) of trypanosomiasis in cattle.

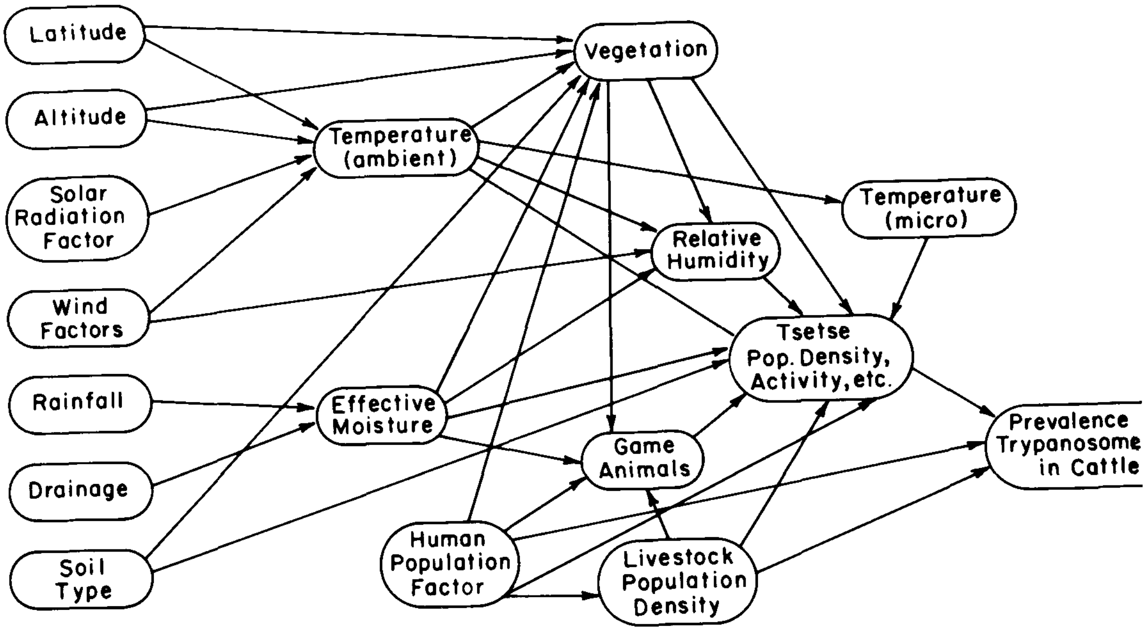

Fig. 1 Path analytical model for African trypanosomiasis.

After the significant determinants were identified, information on specific variables were obtained from Ethiopia for which a five year survey conducted by an Ethiopian and British team (Langridge 1976) provided data about the distribution of tsetse flies and infested areas, and about the prevalence of trypanosomiasis in cattle. The survey was done in Southwestern Ethiopia, which includes 7 administra-regions, each subdivided into smaller areas. The 52 subdivisions or "awrajas" which are somewhat equivalent to counties in the United States constitute the sample size and were referred to as case areas.

Additional cross-regional data were obtained from information provided by the Ethiopian government (Anon 1970-1973, 1977) or by cartographic methods. The cartographic method involved the use of a polar plantimeter to measure surface areas on detailed maps drawn to scale and then converting the measurements to km^2.

All data were coded, programmed and stored on comupter disc files for processing. Since the study involved a large number of predictor variables, stepwise regression-forward selection was utilized to narrow down the field. The computations were done using BMD 02R program, a stepwise regression program which gives combinations of variables that provide the best prediction of the dependent variable using partial correlation coefficients and F ratios. To measure the predictive ability of the variables and evaluate adequacy of fit, the coefficient of determination (R^2) and multiple correlation coefficient (R) were used.

To perform the discriminant analysis and develop a classification function for trypanosomiasis risk the 52 case areas were classified into various risk groups according to the dependent variable: prevalence of trypanosomiasis in cattle. Two examples of classification were considered; each was examined by discriminant analysis using the BMD 7M program (Dixon and Brown 1979) on a Burroughs 6700. The results and misclassification probabilities obtained were interpreted in light of epidemiologic concepts pertinent to trypanosomiasis.

Results

Based on stepwise forward selection the most meaningful predictor subsets were selected. The cutoff point for selecting the best predictor subset was determined by considering the importance of each variable in providing quantitative information that can be used in evaluating disease control alternatives, the R^2 from stepwise regression, and the size of the F ratio. The best predictor equation was:

$$\hat{Y} = 0.43 + 1.388X_{25} + 4.21X_{29} + 4.16X_{26} - 0.0132X_{23} - 0.00468X_{20} + 0.0189X_{18} \qquad (1)$$

The standard errors for each predictor variable in the equation were 1.3026, 1.2345, 0.9689, 0.00634, 0.00276 and 0.01189 respectively. The RSS was minimal, indicating that the unexplained variability (residual) is small. The signs of the regression coefficients were intuitively proper. The correlation between the predictor variables was low except for two cases viz. between X_{19} (cattle population) and X_{20} (sheep and goat population); and X_{26} (indicator variable for Suidae) and X_{25} (indicator variable for Bovidae).

The basic assumptions for regression were examined by graphical analysis. When the residuals ($\hat{\varepsilon}_i$) were plotted against the fitted value ($\hat{Y}_i$), it resulted in a random distribution around the mean value 0, indicating a reasonable approximation of the independence assumption of regression. The assumption of normal distribution, however seemed violated since a histogram of the ε_i was not a very close approximation of a normal distribution although it may have been satisfactory considering the complexity of this problem.

Two cases were considered for the discriminant analysis.

Case I: Critical Prevalence = 5%: The following discriminant functions (D) were obtained when high risk category was set at a prevalence of $\geq$ 5% and low risk at < 5%. For the low risk group:

$$D1 = 5.639X_{12} - 4.105X_{13} - 0.003X_{23} + 15.79X_{28} + 4.888X_{29} - 45.483 \qquad (2)$$

for the high risk group, the discriminant function

$$D2 = 7.470X_{12} - 6.084X_{13} - 0.047X_{23} + 31.539X_{28} + 11.092X_{29} - 73.706 \qquad (3)$$

The discriminating variables were: temperature (average maximum = X_{12} and average minimum = X_{13}), human population density = X_{23} and Glossina types: (morsitans = X_{28}, fusca = X_{29}). A case area is assigned to the group with the largest value of the classification function.

The morsitans tsetse group contributed more than fusca group at this level of categorization. With the aid of these classification functions, each of the 52 case areas were classified as either a high or low risk area. Only one area was misclassified from each group; i.e., 96.2% were correctly classified.

Case II: Critical Prevalence = 10%: The same independent variables were used to classify the 52 case areas into two groups where the high risk group had a prevalence of 10% and the low risk group had a prevalence of 10%. The procedure of variable selection was completed after 3 steps; the variables selected by order of entry were X_7, X_{24}, and X_{29}. The discriminant functions generated were:

$$D1 \text{ (low risk)} = 0.002X_7 + 0.904X_{24} - 0.188X_{29} - 3.168 \qquad (4)$$

$$D2 \text{ (high risk)} = 0.46X_7 + 1.36X_{24} + 6.629X_{29} - 7.66 \qquad (5)$$

It is interesting to note that the fusca group of tsetse (X_{29}) is now important but the morsitans group (X_{28}) did not enter into the discriminant function; this may indicate that the presence of fusca type Glossina may in fact be responsible for the high prevalence. Although 3 useful discriminating variables were obtained, the probability of correct classification achieved (0.846), was lower than the one obtained in the previous example. This probability, however, is still of value considering the very critical level of grouping (10% prevalence).

Discussion

Several analytical steps were required to arrive at the best predictor equation. The selection of the final predictor subset of variables was facilitated by considering the combined inputs of the coefficient of determination (R^2), and of a good understanding of the epidemiology of the disease.

It is significant to note that information from diverse and heterogenous sources such as the ones used in the present study, have provided useful, quantitative information in a mathematical equation. The results indicated that the most productive efforts in the control of trypanosomiasis in cattle in Ethiopia were: a) decreasing the vector population especially fusca group tsetse which are associated with high prevalence (X_{29}), b) limiting/decreasing the game population (X_{25}, X_{26}), c) increasing human population density (X_{23}), and d) increasing sheep and goat population (X_{20}) while holding the cattle population down (X_{19}).

The coefficients associated with the final predictor equation were informative. For example, an increase in feeding on Bovidae resulted in higher infection rates than feeding on Suidae, although the latter were preferred hosts for blood meals for Glossina (Ford 1971). The finding of an association between high trypanosomiasis infection rates and increased blood feeding on Bovidae, has been reported (Jordan 1965). The presence of fusca group tsetse had a higher influence on prevalence than the presence of other groups of Glossina. Thus, the magnitude as well as the direction (sign) of each predictor variable, especially those amenable to external human influence, can be assessed from the final predictor equation.

In the case of the discriminant models, the correct classification probabilities were utilized in deciding whether the discriminant functions obtained were adequate as a classification rule. In this case the functions obtained appeared to be quite effective in discriminating between the two groups.

What would be the epidemiologic consequences, i.e., penalty for misclassification of an area? The consequence of classifying a high risk region into the low risk group might be that the government would implement a resettlement program in that area assuming that it is at low risk. Susceptible cattle and/or humans would enter such an area and epidemics of trypanosomiasis could arise. Resulting mortality and loss of productivity in both cattle and humans would have drastic economic consequences and the zeal of the resettlement program would be reduced; the people who have resettled the area would move out and the region would remain one of high risk. If the area were known to be at high risk before resettlement is initiated, other vector and trypanosomiasis control measures could be implemented along with resettlement and the program could succeed. Mistakenly classifying a low risk as a high risk is not as critical. The result of such a mistake may involve extra investment in disease control expenses and/or the postponement of resettling the region and exploiting the resources. This latter error would not appear to be as costly as the former.

Interpretation of the discriminant function involved evaluating the coefficients of the X_i in the discriminant function of Eq. (2 and 5). A change in the independent variable by one unit changes the discriminant score by the amount of the coefficients of the X_i variable; therefore, the magnitude (size) of the coefficient and its sign are important.

For example, the amount of km^2 of forested land (X_7) has a coefficient of 0.46 in Eq. (5) and 0.002 in Eq. (4). Large areas of forests (X_7) would increase the value of D2 so that the classification into the low risk group is minimized. The direct effect of X_7 on D1 is small; its stronger influence on D2 enhances classification into the high risk group. Therefore, if forested areas were reduced selectively (for example by decreasing forests near farm lands while reforesting other nonutilized land areas), the vector population, especially the fusca group, would decrease. It is known that fusca group tsetse prefer forested areas (Ford 1971); this association is reflected in the discriminant function.

Variable X_{24} (density of humans/cultivated hectare) affects tsetse presence; thus, resettlement programs and maintenance of a high human population density may be vital to keep tsetse from infesting an area. It also increases the production of the land.

The significance of the results of discriminant analysis in light of the objectives of trypanosomiasis control program in Ethiopia was thus demonstrated. The implications of misclassification consequences, and the interpretation of the coefficients of the selected variables were useful in determining rational alternatives of control approaches to the problem faced in Ethiopia. Thus, not only have we gained information of a descriptive and predictive nature for the prevalence of trypanosomiasis in cattle in Ethiopia, but also a useful method of evaluating avenues for the most practical and applicable method of effectively controlling the disease.

REFERENCES

1. Habtemariam, T. 1979. A study of African trypanosomiasis using epidemiologic models; The case of Ethiopia. Ph.D. thesis, Department of Epidemiology and Preventive Medicine, University of California, Davis, CA. 388 p.

2. Ford, J. 1971. The role of the trypanosomiasis in African ecology; a study of the tsetse problem. Clarendon Press, Oxford, England. 568 p.

3. Glover, P.E. 1967. The importance of ecological studies in the control of tsetse flies. Bull. WHO.37: 581-614

4. Wilson, A.J., Paris, J. and Dar, F.K. 1975. Maintenance of a herd of breeding cattle in an area of high trypanosome challenge. Trop. Anim. Hlth. Prod. 7: 63-71

5. Langridge, W.P. 1976. A tsetse and trypanosomiasis survey of Ethiopia. Ministry of Overseas Development, London, England. 98 p.

6. Morrison, D.F. 1976. Multivariate statistical methods. McGraw-Hill Book Co., New York, N.Y. 415 p.

7. Bolch, B.W. and Huang, C.J. 1974. Multivariate statistical methods for business and economics. Prentice-Hall, Englewood Cliffs, N.J. 329 p.

8. Neter, J. and Wasserman, W. 1974. Applied linear statistical models. R.D. Irwin, Inc., Homewood, IL. 842 p.

9. Anon. 1970-1973. Ethiopia. Provincial livestock development studies, Ministry of Agriculture. Addis Ababa, Ethiopia.

10. Anon. 1977. Ethiopia. Animal and Fisheries Resources Development Authority. Addis Ababa, Ethiopia.

11. Dixon, W.J. and Brown, M.B.(ed). 1979. Biomedical Computer Programs P-Series. University of California Press, Berkeley and Los Angeles. 880 p.

12. Jordan, A.M. 1965. The hosts of Glossina as the main factor affecting trypanosome infection rates of tsetse flies in Nigeria. Trans. R. Soc. Trop. Med. Hyg. 59: 423-431.