The Kenya Veterinarian 18(2) (1994) 562-565

- 16.4 -

A PROPOSAL TOWARDS DEVELOPING A SYSTEM OF TEST EFFICIENCY INDICES AIMED AT CLASSIFYING TEST QUALITY WHEN TESTS ARE USED OVER POPULATIONS

SLENNING, B.D.[a]

The function of a test is to help decide the disease status of animals. Historically, test quality has been measured using sensitivity (SEi; test positive animals divided by truly diseased animals), and specificity (SPi; test negative animals over the truly non-diseased animals). Unfortunately, these test quality evaluation measures and their derivatives are valid only when used on an individual animal basis; testing populations yields results not fully encompassed by such standards (Martin et al., 1992; Tyler and Cullor, 1989).

The purpose of this paper is to explore the dynamics of tests applied over populations and to generate new metrics that might better reflect the situations. Specific questions are: (1) Should SEi or SPi be given priority in population testing situations? (2) What is the impact of population size and integer-level calculations on efficiency? (3) How does disease prevalence within or between groups affect efficiency? (4) Can we predict performance without knowing true prevalence of disease in a population?

MATERIALS AND METHODS

An electronic spreadsheet-based model (Lotus Corp., 1992) was developed to create a series of scenarios for population-level testing. The model assumes the disease in question to be severe, so that if a herd test reveals one positive animal, the herd is considered positive. The spreadsheet allows the user to select population prevalence (P-P) scenarios consisting of a range of sizes for four tested groups over five levels of disease prevalence to select values for test sensitivity and specificity. The model then calculates single group and multiple group values for true positives (TP), true negatives (TN), and false positives (FP) and negatives (FN), and generates predictive values for positive and negative tests (PVPT and PVNT), and the ratio of correct to incorrect classifications (T:F). It summarizes by developing population-level sensitivity and specificity (SEp and SPp, respectively), predictive values (PVPTp and PVNTp), three measures of prevalence: theoretical ('target' prevalence), actual ('true' prevalence - number of diseased herds/total numbers of herds), and test ('apparent' prevalence - number of +ve tested herds/total number of herds)), and ratio of correct to incorrect classifications (T:Fp).

P-P scenario choice was made thus. The smallest testable population is a group of two. Hence, all scenarios use a size of two as the smallest of four populations. The magnitude of population increase was chosen to vary from 20 to 100, with roughly equal steps in between (i.e., for the 2-20 size range, the steps were 2, 8, 14, and 20). Disease prevalence in most programs was assumed $<=50\%$; hence a range of 0%, 10%, 20%, 30%, 40% was used (sensitivity analyses required additional ranges of 0-12%, and 0-100%). A P-P scenario, for the purposes of this model, is signified by a size range (i.e., 2/8/14/20) and a prevalence range (i.e., 0/.1/.2/.3/.4). The model generates efficiency values for each combination of size and prevalence ($n=20$) and summarizes overall efficiency by evaluating performance over the entire scenario.

Scenario population-level performances are compared graphically for trends as the input variables change. Quality of scenario is determined by patterns of stability in the output indices. Such analysis addresses questions (1), (2), and (3) above. Question (4), predicting outcomes (true prevalence and PVPTp), is accomplished through regression analysis, using linear regression to select the best multiple regression model based on Mallow's CP and adjusted r-square (Analytical Software, 1990).

[a]Population Medicine Program, Department of Food Animal and Equine Medicine, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, 27606, USA.

RESULTS AND DISCUSSION

The choice of a severe disease first warrants discussion. Our perspective of disease importance determines the working definition of what constitutes a 'diseased' herd. For extremely damaging diseases such as tuberculosis or Newcastle's disease, a finding of one infected animal results in the animal's entire herd or region being declared infected. This is the assumption used in this paper. It should also be noted, however, that we also need to deal with diseases where we accept levels of infection below a certain threshold. For instance, *Staphylococcus aureus* mastitis in dairy cows is often allowed to remain at low levels. Such a situation calls for a different set of assumptions and analyses than those presented here.

Should SEi or SPi be given priority in population testing situations?

Figures 1 and 2 display the results of applying a test over a P-P scenario and changing SEi and SPi, respectively. The P-P scenario is 2/8/14/20-0/.1/.2/.3/.4. For this scenario, SEi has essentially no impact on ability to identify infected herds (SEp stays at 100% throughout).

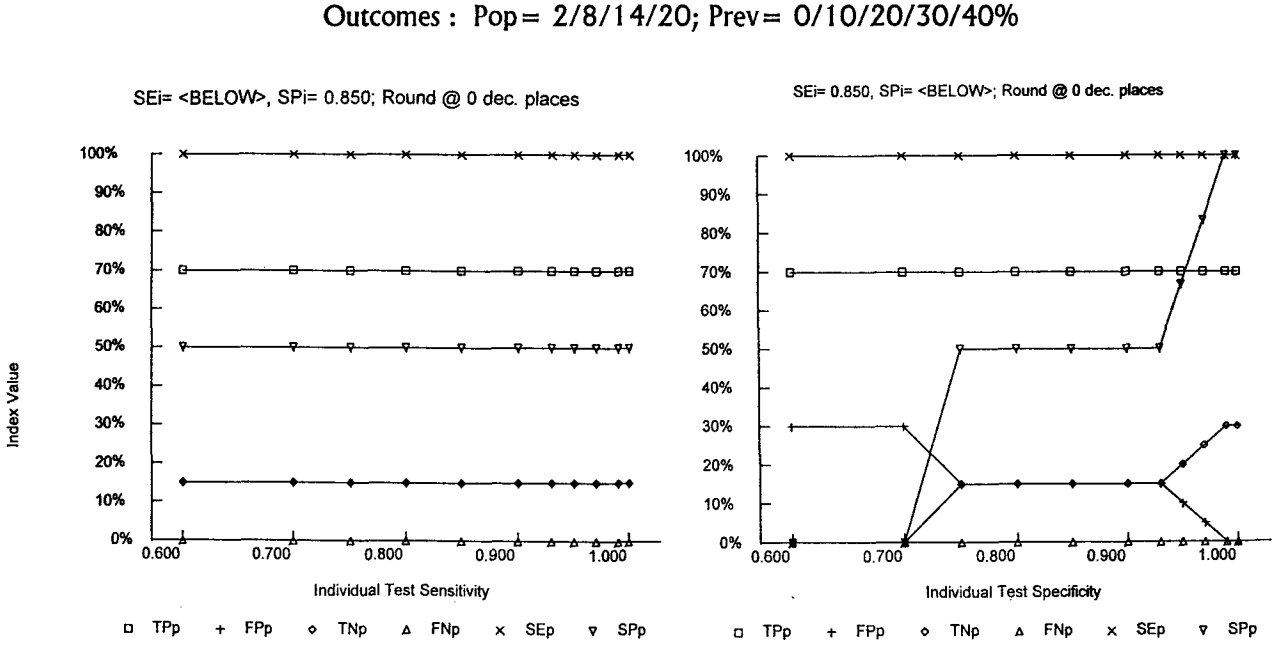

Fig.1. Impact of changing SEi on population-level test performances.

Fig. 2. Impact of changing SPi on population-level performance

Altering SPi, however, gives a different result. While the SEp still remains constant at 100% (i.e., FN=0%), the test's ability to avoid FP is severely impacted. At SPi of 0.70 or lower, SPp is 0.0% so the test cannot identify truly non-infected herds. All herds in this situation test positive. At SPi of 0.75, the SPp jumps to 50%, where it remains until SPi goes over 0.92. It quickly approaches 1.0 thereafter. Hence, test performance over a population is greatly impacted by SPi, but not at all by reasonable changes in SEi. This result was seen over many different P-P scenarios.

What is the impact of population size and integer-level calculations on efficiency?

One deficiency of classical applications of SEi and SPi to populations is that rounding assumes we can have decimal portions of animals. We cannot; we have whole animals. As a result, with small population sizes the estimates of performance make 'step' changes as opposed to smooth curvilinear changes.

Figure 2 shows this effect as well. Note how SPp makes jumps of 50%. This is the effect of small numbers and of rounding at the integer level. This is what we see in reality, and is the basis of the remaining analyses.

How does disease prevalence within and between groups affect efficiency?

The impact of disease prevalence on test performance is well known; as prevalence increases, so does PVPT (i.e., FP goes down), and as prevalence goes down, so does PVNT (i.e., FN decreases) (Tyler and Cullor, 1989). Over populations, however, the relationship is not so clear. If SPp makes step changes (Fig. 2), predictive values will change similarly.

Figure 3 shows how, at 0% prevalence, the test identifies half the herds correctly (PVNTp=50%), yet at 10% prevalence the test identifies all herds correctly (PVPTp= PVNTp= 100%). This works out, in part, because of our definition of a 'diseased herd' (i.e., one infected animal = 'disease'), the nature of not allowing decimal levels of animals to exist (causing actual prevalence to differ from theoretical prevalence), and the punctuated improvement in SPp as theoretical disease prevalence increases.

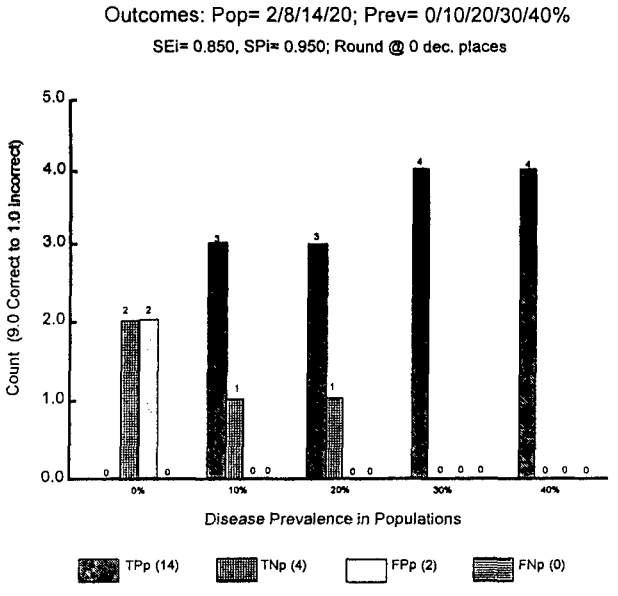

Fig. 3. Impact of prevalence on performance. As prevalence increases performance makes punctuated improvements.

Can we predict performance without knowing true prevalence of disease in a population?

The above discussion of SEi and SPp, etc., is of little value in practice unless it allows us to make better decisions. One way for this to occur is if the summary figures allow us to calculate true prevalence or population-level predictive values.

Multiple P-P scenario simulations were carried out where SPi was varied from 0.90 to 0.99, and prevalence varied from 0% to 100%. A simple multiple regression was applied using population size, test prevalence, individual-based test specificity, with the addition of a population size transformation and a population size test prevalence interaction term.

From 189 individual simulations, the regression significantly ($P < 0.01$) fitted the actual PVPTp, explaining 76% of the variability in actual PVPTp. At lower PVPTp (about 65%) the regression lost its ability to fit the actual PVPTp line. The same occurred at extremely high PVPTp (1.0). If the prevalence was allowed to vary only by 20% (e.g., from 20% to 40%), the variability explained could reach the upper 80% range.

CONCLUSIONS

This paper demonstrates initial explorations into establishing population-level metrics for evaluating test efficiency. Simple numerical methods applied consistently to population-prevalence scenarios allowed understanding of test outcome dynamics and prediction of test performances and true prevalences. With improvements to these methods, the understanding of and ability to work with disease control programs over populations can be standardized and underlying efficiencies and weaknesses can be compared across tests and across populations. Without such a set of tools we will be hindered in our efforts to improve the well-being of our animals and our client farms.

REFERENCES

Analytical Software, 1990. Statistix User's Manual. Analytical Software, Saint Paul, MN, USA, 280 pp.

Lotus Development Corp., 1992. Lotus 1-2-3, v. 2.4. Lotus Development Corp. Cambridge, MA, USA.

Martin, S.W., Shoukri M., Thorburn, M.A., 1992. Evaluating the health status of herds based on tests applied to individuals. Prev. Vet. Med., 14:33-43.

Tyler, J.W., Cullor, J.S., 1989. Titers, tests, and truisms: rational interpretation of diagnostic serologic testing. J. Am. Vet. Med. Assoc., 194:1550-1558.