European Journal of Plant Pathology **102:** 847–854, 1996.

Calibration and verification of risk algorithms using logistic regression

Jonathan Yuen[1], Eva Twengström[2] and Roland Sigvald[2]
[1]*Department of Cancer Epidemiology, Uppsala University, University Hospital, S-751 85 Uppsala, Sweden (Fax: 46 18 503431); and Department of Plant Pathology, Swedish University of Agricultural Sciences, PO Box 7044, S-750 07 Uppsala, Sweden;* [2]*Research Information Centre, Swedish University of Agricultural Sciences, PO Box 7044, S-750 07 Uppsala, Sweden (Fax: 46 18 672890)*

Accepted 27 June 1996

Abstract

The use of logistic regression is proposed as a method of verifying and calibrating disease risk algorithms. The logistic regression model calculates the log of the odds of a binary outcome as a function of a linear combination of predictors. The resulting model assumes a multiplicative (relative) relationship between the different risk factors. Computer programs for performing logistic regression produce both estimates and standard errors, thus permitting the evaluation of the importance of different predictive variables. The use of receiver operating characteristic (ROC) curves is also proposed as a means of comparing different algorithms. An example is presented using data on Sclerotinia stem rot in oil seed rape, caused by *Sclerotinia sclerotiorum.*

Introduction

Risk algorithms have been designed and used to predict whether future disease levels in individual fields will exceed a certain level in the future. If the predicted risk is sufficiently high, then control measures, often in the form of protective sprays, are recommended. Traditional risk algorithms include those proposed for late blight of potato caused by *Phytophthora infestans* (Krause et al., 1975), Alternaria leaf blight of carrots caused by *Alternaria dauci* (Gillespie and Sutton, 1975), and apple scab caused by *Venturia inaequalis* (Mills, 1944). In these systems, risk 'points' (calculated in lookup tables as a function of weather and other risk factors) are summed and sprays are then recommended after a threshold is exceeded.

The origin of these risk algorithms is not always clear, though they are undoubtedly based on extensive experience and subjective judgement. The unclear origins can create problems if these algorithms (in unmodified form) are to be used in geographic areas with different climatic conditions or with other host plant cultivars with different levels of resistance. The relatively subjective nature of the development process also makes it difficult to develop new algorithms for other pathosystems in the absence of extensive experience.

An objective method for calibrating and verifying these algorithms is needed if functional ones are to be used outside of the area of development, or if new ones are to be developed without extensive experience. In this paper, we describe a method of evaluating and calibrating risk algorithms based on logistic regression, and the graphical comparison of different risk algorithms via receiver operating characteristics (ROC) curves.

Materials and methods

Original algorithm

In Sweden, *Sclerotinia sclerotiorum* (Lib.) deBary generally infects oil seed rape during flowering, causing Sclerotinia stem rot. Adequate control of the disease can generally be obtained by application of a single fungicide spray during flowering. This fungicide application is not required every year, and considerable savings can be obtained by only applying fungicide when necessary.

A disease prediction algorithm for Sclerotinia stem rot in oil seed rape using a risk point sum was described by Twengström and Sigvald (1993) (Table 1). If the total number of points was less than 60, sprays were not recommended, and between 60 and 100 was an intermediate, uncertain zone. A total of 100 or more points indicated high risk for *Sclerotinia sclerotiorum* infection and fungicide application was recommended.

While disease level at harvest is a continuous variable, the outcome from this algorithm represents a need to apply a fungicide spray. This is a binary variable corresponding to a final disease level over 20% (a spray was needed) or less than or equal to 20% (a spray was not needed). The decision threshold was determined by the point where the cost of fungicide applications were offset by the increase in yield, here referred to as the the economic action threshold (Twengström and Sigvald, 1993).

Data collection

Rapeseed growers in the provinces of Uppland and Västmanland in eastern parts of Central Sweden were asked to leave a 20m by 50m plot unsprayed during the years 1988 to 1991. Past cropping history and present cultural practices as well as current and forecasted weather predictions were also collected so that a disease risk forecast could be calculated for each field. At the end of the growing season, 200 plants were randomly selected from the untreated plots in these fields and used to calculate disease incidence. A total of 267 fields were available for analysis. The original disease prediction algorithm was used to calculate the point score, and a particular field was judged to be in need of a spray if the final disease incidence exceeded the economic action threshold.

Logistic regression

For the purposes of this study, fields were divided into two groups (over 20% and 20% or less diseased plants, corresponding to the economic action threshold used in developing the original algorithm), yielding a binary outcome variable. Independent variables were the original weather and cropping history data from the respective fields.

A comprehensive description of logistic regression would be difficult to give in this space, but generally stated it calculates the probability of a given binary outcome as a function of a set explanatory variables (McCullagh and Nelder, 1989). If we denote the true risk or probability of the 'outcome' as P, the model assumes that the logarithm of the odds of P (odds(P) = P/(1 − P)) is a linear function of the explanatory variables. Note that the logarithm of the odds of (P) is logit(P). In this particular situation, the binary outcome is the probability that the final disease level in a given type of field (defined by a unique combination of the covariates) exceeds the economic action threshold. It could also be considered the probability of the need to apply fungicides in a given field. The model fits the logit of this probability as a linear function of the covariates (such as cropping history and weather data) for that field. Differences between the observed outcome and the predicted outcome (P) are accomodated by assuming the predicted values follow a binomial distribution. For further details regarding logistic regression, it is suggested that the reader consult Hosmer and Lemeshow (1989), or a reference work on general linear models (McCullagh and Nelder, 1989).

Normal least squares methods cannot be used to calculate the regression coefficients in logistic regression, and maximum likelihood methods are required. Both GLIM with binomial error and logistic link (The Numerical Algorithms Group Ltd, Wilkinson House, Jordan Hill Road, Oxford, United Kingdon OX2 8DR) and the SAS Procedures GENMOD and LOGISTIC (SAS Institute Inc., SAS Circle, Box 8000, Cary, NC 27512–8000) were used to fit these models with identical results. All these methods yield, in addition to the parameter estimates, the variance and covariance matrix of the parameter estimates. The standard deviation of the parameters (the square root of the variance) was used to calculate 95% confidence intervals for the estimates. Candidate models were compared using their deviance (−2 times log likelihood). The difference in deviance between two models was tested against a χ^2 value (the number of degrees of freedom corresponding to the difference in degrees of freedom between the two models), i.e. a likelihood ratio test (LRT) (McCullagh and Nelder, 1989).

Risk factors were evaluated in two ways. In the first, the ability of a risk factor alone to predict the risk that a field became diseased was tested. This consists of testing the deviance reduction attributed to a variable when it is first entered into the model. In the other method, the predictive ability of a factor was tested by removing it from a complete model with all explanatory variables. The latter method, referred to as a Type 3 analysis in Genmod (SAS Institute, 1993), duplicates the analysis of deviance entry that a variable would have if it was the last variable entered. If

Table 1. Risk factors with points from original prediction algorithm, with likelihood ratio test (LRT) for that factor entered into a model first, and for that factor entered into a model last

Risk factor	Possible answers	Points	LRT Factor first	LRT Factor last
Number of years since last oil-seed crop	More than 6 years	0	1.097	8.819
	5–6 years	5	2 df[a]	2 df[a]
	3–4 years	15		
	1–2 years	20		
Level of infection in last oil-seed crop	None	0	18.275	15.819
	Low (1–10%)	10	5 df	5 df
	Moderate (11–30%)	20		
	High (31–100%)	30		
	Don't know (low risk)	10		
	Don't know (high risk)	20		
Have peas been grown in the field during the last five years Foliage density (including weeds) 0.5 m above ground	No	0	0.108	0.054
	Yes	5	1 df	1 df
	Thin	0	12.387	12.818
	Normal	15	2 df	2 df
	Heavy	30		
Rain in June	Less than normal	0	40.62	1.438
	Normal (35–55 mm)	10	2 df	2 df
	More than normal	20		
Rain the last two weeks	Less than 10 mm	0	17.638	1.637
	10–30 mm	15	1 df[b]	1 df[b]
	More than 30 mm	25		
Weather forecast	High pressure	0	12.612	9.601
	Variable	5	2 df	2 df
	Low pressure	10		
Regional risk value for apothecia development (per 100 sclerotia)	0–5	0	51.083	19.082
	6–10	5	2 df[c]	2 df[c]
	11–20	15		
	21–100	25		

[a] The groups labelled 1–2 years and 3–4 years were combined to form one group in the logistic regression analysis.

[b] The groups less than 10 mm rain and 10–30 mm rain were combined to form one group in the logistic regression analysis.

[c] The groups labelled 11–20 apothecia and 21–100 apothecia were combined to form one group in the logistic regression analysis.

none of the independent variables are correlated, the two LRT's would be identical. Since the independent variables here are related, any complete analysis of deviance tables would be dependent on the order of the variables. This method (examining the importance of a variable if entered first and last) duplicates the essential portion of the analysis of deviance table in a manner that is not dependent on the order of the variables. A complete analysis of deviance table was generated for the final model, where the reduction in the residual deviance was calculated for each variable as it was added to the model.

Both the original and the modified algorithm produce a range of values, and since neither is a perfect predictor, any decision threshold based on these algorithms (such as 'spray if the point accumulation is above a given level') yields one of four possible situations (Table 2). The cells A and D represent the number of correct decisions, whereas B and C represent the incorrect decisions. If a total of A+C fields had a final

Table 2. Calculation of true and false positive error rates for a predictor of Sclerotinia stem rot in oil-seed rape

	Disease exceeds 20%	Disease less than 20%
Spray	A	B
Don't spray	C	D
Total	A+C	B+D
	A/(A+C) True positive	B/(B+D) False positive

disease level over 20%, and our algorithm predicted this would occur in A fields, then the true positive rate (TP) would be A/(A+C). Likewise, if (B+D) fields had a final disease level less than 20%, and our algorithm recommended a spray for B of these fields, then the false positive rate (FP) would be B/(B+D). The TP and FP rates can be affected by changing the decision threshold. For example, recommending a spray with a lower point accumulation would increase the TP rate, but may also increase the FP rate. An ideal algorithm would maximize the TP rate and minimize the FP rate.

The relationship between TP and FP for the risk prediction algorithms was examined graphically by plotting receiver operating characteristic (ROC) curves (Metz, 1978). These curves plot the TP as a function of the FP at all possible decision thresholds. In an ROC curve using our example, the origin of the graph represents the decision 'not-to-spray' for all fields. This decision yields no false positives (i.e. a recommendation to spray in the absence of the need to spray: the quantity B in Table 2 is 0) but captures no true positives (recommending a spray for those fields that truly require one: the quantity A in Table 2 is also 0). The upper right corner would recommend spraying all fields, thus detecting all fields that truly require a spray (C = 0, TP rate = 1) but also recommending a spray for all fields that do not require them (D = 0, FP = 1). An efficient algorithm would yield a curve 'pushed to the upper left corner'.

Results

The risk factors proposed in the original algorithm varied in their ability to predict final disease in the fields tested (Table 1). Some factors, such as level of infection in the last rapeseed crop, were of importance irregardless of which other risk factors were included in the model. Other factors, such as whether peas had been grown in the field during the last five years, were not able to predict the need to spray alone or in combination with other risk factors. A third group were factors that could be significant by themselves, but which were not significant after other risk factors were included in the model. The number of years since the last oil-seed crop was initially not important (χ^2 = 1.097, 2 df, when entered first), but gained in importance with the addition of other variables (χ^2 = 8.819, 2 df, when entered last). This was due to the regional risk value (data not shown).

A subset of the most important risk factors were tested in a reduced model that included number of years since last oil-seed crop, level of infection in last oil-seed crop, foliage density, rain in June, and weather forecast. Peas as a previous crop was eliminated since it was unrelated to the risk of stem rot. The regional risk value was also left out of the model in order to test the performance of an algorithm without this factor, since it is expensive to measure, and cannot be easily used without access to sclerotia and proper training. Rain within the last two weeks and rain in June were highly correlated with each other. Rain in June was a better predictor (Table 1), and was retained. Analysis of deviance (Table 3) showed that most variables in this model were important. The number of years since the last crop was not statistically significant in this particular model since the regional risk value was not included. The parameter estimates resulting from the regression (the natural logarithm of the odds ratio) and their standard errors are presented in Table 4. Exponentiating the parameter estimate yields the odds ratio, which can be interpreted as the relative risk. Both values are useful, since the parameter estimates have an additive relationship, and can thus be summed in a manner similar to the original risk algorithm, whereas the relative risk values themselves are easier to interpret.

The original prediction algorithm varied in its ability to predict whether disease incidence would exceed the economic action threshold (Figure 1). The reduced, calibrated model with only five risk factors was similar in its ability to predict the need to spray (Figure 2). Both the original and the reduced model had a wide range of values for fields that developed little or no disease. This is somewhat underestimated in the Figures 1 and 2 since several points near the Y axis are actually several observations. A simultaneous comparison of the ROC curves from the original and calibrated risk algorithm also shows relatively small differences between the two algorithms in their ability to predict the need for fungicide treatments (Figure 3).

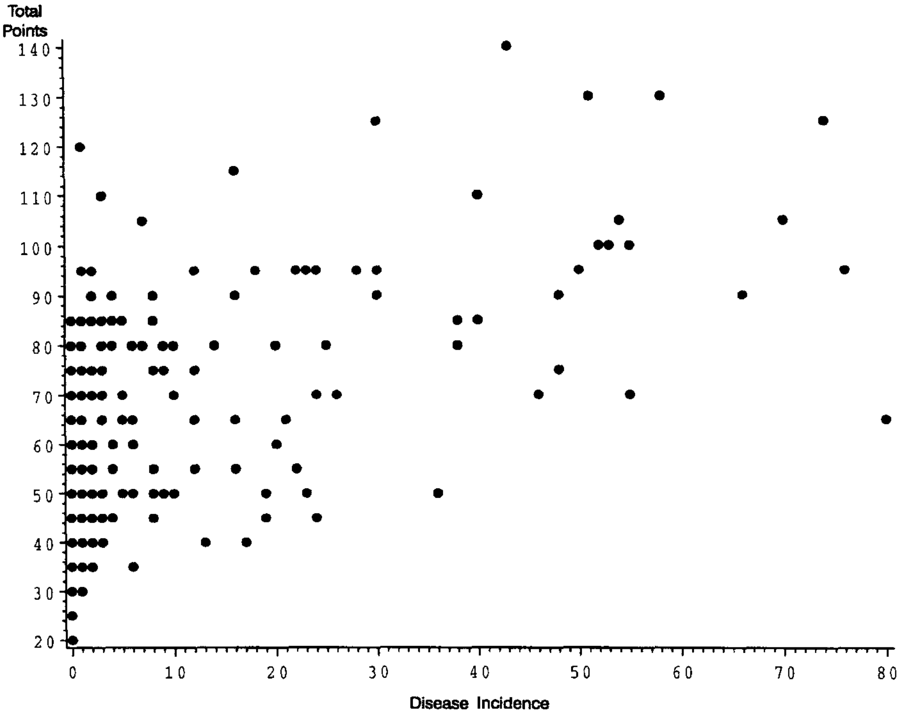

Figure 1. Accumulated risk points (original scale) as a function of disease incidence in 267 oil seed rape fields from 1988 to 1991.

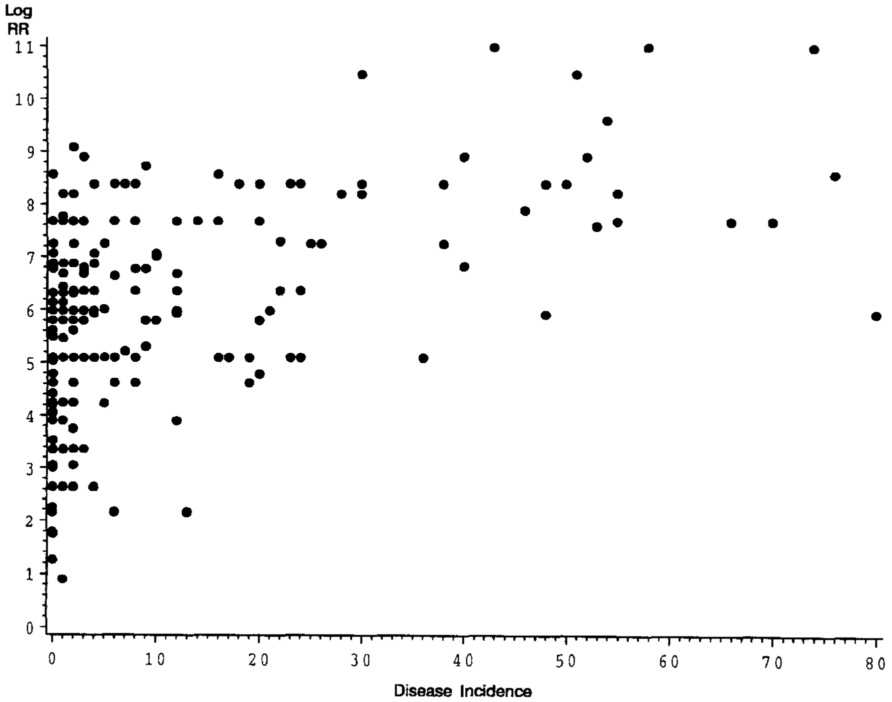

Figure 2. Log relative risk from a reduced model (number of years since last oil-seed crop, level of infection in last oil-seed crop, foliage density, rain in June and weather forecast) as a function of final disease assessment from 267 oil seed rape fields from 1988 to 1991.

Table 3. Analysis of deviance showing relative importance of of adding various factors in a reduced model predicting 'need to spray' (based on final disease incidence over 20%)

Added Factor[a]	Deviance[b]	DF[c]	χ^2	Pr> χ^2
Intercept	211.1804	0	.	.
Number of years since last oil-seed crop	210.0831	2	1.0974	0.5777
Level of infection in last oil-seed crop	191.1924	5	18.8907	0.0020
Foliage density	180.7123	2	10.4800	0.0053
Rain in June	148.9162	2	31.7962	0.0000
Weather forecast	138.8070	2	10.1091	0.0064

[a] Each factor added to the model listed in the previous line.

[b] Residual deviance of model.

[c] Degrees of freedom added to model due to addition of the factor.

Table 4. Risk factors, natural logarithm of relative risk, and standard error for a reduced model to predict 'need to spray' (based on final disease incidence over 20%)

Risk Factor	Possible response	$\log_e$(RR)	SE	Relative risk
Number of years since last oil-seed crop	More than 6 years	Reference	–	1
	5–6 years	0.6074	0.5414	1.84
	1–4 years	−0.8506	0.8828	0.43
Level of infection in last oil seed crop	None	Reference	–	1
	Low	1.4497	1.2269 4.26	
	Medium	1.7098	1.4527	5.53
	High	3.9872	1.4034	53.90
	Don't know (low risk)	1.9424	1.2334	6.98
	Don't know (high risk)	1.078	1.1744	2.94
Foliage density	Thin	Reference	–	1
	Normal	1.8974	0.88	6.67
	Dense	2.4075	0.9854	11.11
Rain in June	Less than normal	Reference	–	1
	Normal	1.1920	0.6791	3.293662
	More than normal	2.7092	0.5976	15.01726
Weather forecast	High Pressure	Reference	–	1
	Variable	2.7749	1.1817	16.03702
	Low pressure	1.17711	1.5519	5.88

Discussion

The use of logistic regression allows for the development and calibration of risk algorithms from empirical data. Using logistic regression, the importance of various risk factors can be evaluated in a systematic manner. The increased risk resulting from the presence of a given risk factor can be readily quantified, along with an indication of the amount of variability associated with this estimate. Use of relative risk also implies a multiplicative relationship between the different risk factors. Thus if factor A increases the risk by a factor of 2 and factor B by a factor of 1.5, the presence of both A and B would increase the risk by a factor of 2 × 1.5 or 3. An iterative method can be used to further refine these risk estimates. As additional data are accumulated (for example, after the current growing season), the risk estimates can be recalculated and the resulting new algorithm tested the next growing season.

The use of ROC curves is a method of comparing risk algorithms that does not rely on the algorithms having the same scale. An added advantage is that it allows for flexibility on the part of the decision maker. Variable decision thresholds, with varying TP and FP rates, can reflect the risk attitudes (utility functions)

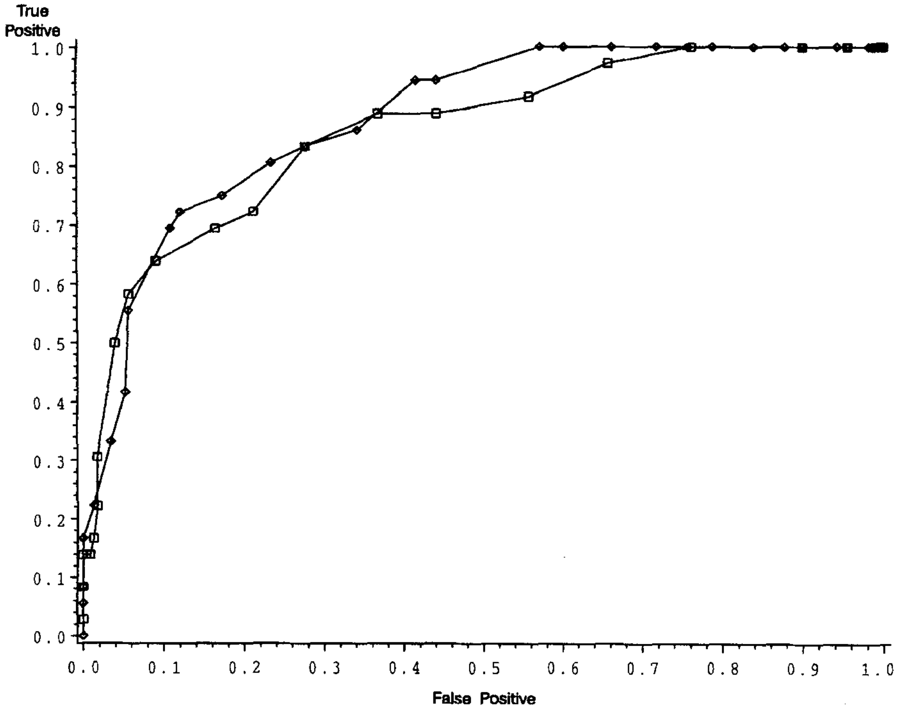

Figure 3. Receiver operator characteristic curves from the original disease forecast algorithm (—□—) and recalibrated algorithm (—◆—). The space between each marker represents a change of 20 points for the original algorithm and a change of 1 (measured as $\log_e$RR) for the recalibrated algorithm.

of the decision maker (Teng and Yuen, 1991). Thus, a risk averse decision maker may set his criteria level further to the right on Figure 3, and spray his fields with a lower 'point accumulation', when compared to a decision maker more willing to take risks. He could thereby increase his true positive rate, but at the cost of increasing his false positive rate. The advantage of the ROC curves is that the rate of both kinds of errors (applying an unneeded spray, and missing a needed spray, in our example) can be estimated. Given these error rates and the relative costs of both kinds of error, the decision maker can determine a critical value for his decision threshold, that reflects his attitudes toward risk.

One should be cautious in using these estimates for risk prediction in rapeseed fields without testing with additional data. Since the data used to compare the algorithms were the data used to calibrate the risk estimates, the results we report here are of limited value for forecasting purposes. The relationship between the number of years since the last crop, the other independent variables, and the development of disease needs to be explored further. Certainly additional years of data, along with verification with an independent data set would be required to develop risk algorithms with wide applicability. This is currently being done with a larger data set by one of the authors (E.T.). Despite the limited data set, the risk estimates reflect the current knowledge of the biology of this particular disease, with increased risks from higher levels of infection in the last crop, a more dense crop canopy, increased precipitation, and a weather prediction indicating precipitation. The factor for number of years since the last crop was not important in this model, though it is important in the presence of the regional risk value (Table 1). The regional risk value is a bioassay of the sum of all factors that affect apothecial development, and this quality of information may be required in order for the factor for the number of years since the last crop to be important.

Logistic regression has been used in plant pathology primarily to analyze disease incidence or other outcome data that is a proportion, such as the proportion of spores that have germinated (Fitzell et al., 1984; Schuh et al., 1987; Dodd et al., 1991; Schuh, 1992; Johnson et al., 1994). In such analyses, the independent variables are similar, but the outcome of such analyses represents disease incidence or the proportion of spores which have germinated. A related method, Poisson regression, has been used to characterize leatherjacket (*Tipula* spp) populations (McCracken et al., 1994). In this case, the dependent variable is the number of larvae in a sample. Little work has been done using the need to apply some control measure as the outcome variable.

Logistic reqression is a method widely used in human epidemiology to quantify risk factors in retrospective studies, commonly called case/control studies (Hosmer and Lemeshow, 1989). The method allows for sampling to take place after the diseased individuals have been identified, so that the sampling proportion of the healthy and diseased population is not the same, or is even known. Typically, close to 100% of the diseased individuals (cases) are sampled, but for rare diseases, less than 1% of the healthy population is sampled. As long as the exposures of interest do not differentially bias the sampling procedures for the cases and the controls, these sampling fractions cancel out each other in the calculation of the odds ratio. One extension of the method allows for several response categories (Agresti, 1990) which could be useful if the economic action threshold (in this case the 20% final disease level) can vary. The different response categories can take several forms (nominal, ordinal, or interval variables), which in combination with the postulated relationship between the outcomes and the independent variables, affect the structure of the final model (McCullagh and Nelder, 1989). Attempts to use these methods did not give satisfactory results with these data.

The development of risk assessment algorithms needs to be guided by the required specificity of the final algorithms. If data from a wide variety of cropping situations (variation in cultivars, cultural practices, climate etc.) is used to develop these algorithms, the result may have wider applicability at a loss of specificity. Likewise, restricting the variation in the underlying data may not permit the discovery of important factors, merely because they are always present (or absent), though this may not necessarily affect the validity of the resulting algorithm in that particular setting.

References

Agresti A (1990) Categorical Data Analysis. John Wiley and Sons, New York, 558 pp

Dodd JC, Estrada AB, Matcham J, Jeffries P and Jeger MJ (1991) The effect of climatic factors on *Colletotrichum gloesporioides*, causal agent of mango anthracnose, in the Phillipines. Plant Pathology 40: 568–575

Fitzell RD, Peak CM and Darnell RE (1984) A model for estimating infection levels of anthracnose disease of mango. Annals of Applied Biology 104: 451–458

Gillespie TJ and Sutton JC (1979) A predictive scheme for timing fungicide applications to control alternaria leaf blight in carrots. Canadian Journal of Plant Pathology 1: 95–99

Hosmer DW and Lemeshow S (1989) Applied Logistic Regression. John Wiley and Sons, New York, 307 pp

Johnson DA, Alldredge JR and Allen JR (1994) Weather and downy mildew epidemics of hop in Washington state. Phytopathology 84: 524–527

Krause RA, Massie LB and Hyre RA (1975) BLITECAST: A computerized forecast of potato late blight. Plant Disease Reporter 59: 95–98

McCracken DM, Foster GN and Kelly A (1994) Factors affecting the size of leatherjacket (Diptera: Tipulidae) populations in pastures in the west of Scotland. Applied Soil Ecology 2: 203–213

McCullagh P and Nelder JA (1989) Generalized Linear Models, 2nd Edition. Chapman and Hall, London, 511 pp

Metz CE (1978) Basic Principles of ROC Analysis. Seminars in Nuclear Medicine 8: 283–298

Mills WD (1944) Efficient use of sulfur dusts and sprays during rain to control apple scab. New York Agricultural Experiment Station Ithaca Extension Bulletin 630, 4 pp

SAS Institute., 1993. SAS Technical Report P-243, SAS/STAT Software: The GENMOD Procedure, Release 6.09, Cary, NC: SAS Institute Inc. 88 pp

Schuh W (1992) Effect of pod development stage, temperature and pod wetness duration on the incidence of purple seed stain of soybeans. Phytopathology 82: 446–451

Schuh W, Jeger MJ and Frederiksen RA (1987) The influence of soil temperature, soil moisture, soil texture and inoculum density on the incidence of sorghum downy mildew. Phytopathology 77: 125–128

Teng PS and Yuen JE (1991) Epidemic models: Lessons from plant pathology. In: Levin MA and Strauss HS (eds) Risk Assessment in Genetic Engineering (pp. 272–296) McGraw Hill, Inc., New York

Twengström E and Sigvald R (1993) Forecasting Sclerotinia stem rot using meteorological and field specific data. SP-Report, Danish Institute of Plant and Soil Science, No 7. Proceedings of the Workshop on Computer-based DDS on Crop Protection, Parma, Italy, 23–26 November 1993