

# A Statistical Comparison of the Blossom Blight Forecasts of *MARYBLYT* and *Cougarblight* with Receiver Operating Characteristic Curve Analysis

M. M. Dewdney, A. R. Biggs, and W. W. Turechek

First and third authors: Department of Plant Pathology, Cornell University, Geneva, NY 14456; and second author: West Virginia University, Tree Fruit Research and Education Center, P.O. Box 609, Kearneysville, WV 25430.

Current address of W. W. Turechek: U.S. Department of Agriculture–Agricultural Research Service, Subtropical Plant Pathology, Fort Pierce, FL 34945.

Accepted for publication 4 April 2007.

## ABSTRACT

Dewdney, M. M., Biggs, A. R., and Turechek W. W. 2007. A statistical comparison of the blossom blight forecasts of *MARYBLYT* and *Cougarblight* with receiver operating characteristic curve analysis. *Phytopathology* 97:1164–1176.

Blossom blight forecasting is an important aspect of fire blight, caused by *Erwinia amylovora*, management for both apple and pear. A comparison of the forecast accuracy of two common fire blight forecasters, *MARYBLYT* and *Cougarblight*, was performed with receiver operating characteristic (ROC) curve analysis and 243 data sets. The rain threshold of *Cougarblight* was analyzed as a separate model termed *Cougarblight and rain*. Data were used as a whole and then grouped into geographic regions and cultivar susceptibilities. Frequency distributions of cases and controls, orchards or regions (depending on the data set), with and

without observed disease, respectively, in all data sets overlapped. *MARYBLYT*, *Cougarblight*, and *Cougarblight and rain* all predicted blossom blight infection better than chance ( $P = 0.05$ ). It was found that the blossom blight forecasters performed equivalently in the geographic regions of the east and west coasts of North America and moderately susceptible cultivars based on the 95% confidence intervals and pairwise contrasts of the area under the ROC curve. Significant differences ( $P < 0.05$ ) between the forecasts of *Cougarblight* and *MARYBLYT* were found with pairwise contrasts in the England and very susceptible cultivar data sets. Youden's index was used to determine the optimal cutpoint of both forecasters. The greatest sensitivity and specificity for *MARYBLYT* coincided with the use of the highest risk threshold for predictions of infection; with *Cougarblight*, there was no clear single risk threshold across all data sets.

Fire blight, caused by the bacterium *Erwinia amylovora* (Burr.), has troubled apple (*Malus ×domestica*) and pear (*Pyrus communis*) growers for over 200 years. It has become more problematic to apple growers over the last 50 years because of increasing acreage planted with high-density production systems that use highly susceptible rootstocks, often in combination with equally susceptible cultivars (18). Many researchers have investigated ways to manage and control disease outbreaks; however, due to the sporadic nature of fire blight, this has proven to be extremely difficult (34). The fire blight disease cycle includes several phases: canker blight, blossom blight, shoot blight, trauma blight, and rootstock blight (29). To prevent disease spread later in the season, it is crucial to control blossom blight early, and one of the few ways to control blossom blight is by the use of antibiotic sprays. Until recently, the ability to predict the onset of blossom blight epidemics accurately and reliably for the timely application of antibiotics has been one of the most limiting factors in improving overall disease management.

Of several fire blight prediction systems, two of the best-known are *MARYBLYT* and *Cougarblight*. Both of these forecasters use weather and apple or pear phenology to generate infection predictions, but they differ in how risk values are calculated. The DOS-based computer program *MARYBLYT*, developed in Maryland and

West Virginia by Steiner and Lightner, includes several algorithms for the different phases of fire blight (canker blight, blossom blight, shoot blight, and trauma blight) and predicts both infection events and symptom development (29). We have concentrated our efforts on apple blossom blight forecasts and, consequently, have listed only the rules for that phase. Blossom infection is considered to be imminent (in prediction mode) when the minimum values of all four of the following variables are met: (i) open flowers with stigmas and petals intact; (ii) epiphytic infection potential (EIP)  $\geq 100\%$ , which is 110 degree-hours at base 18.3°C accumulated in the last 44.4 degree-days at base 4.4°C; (iii) precipitation event of either dew,  $\geq 0.25$  mm on the current day, or previous day rainfall ( $\geq 2.5$  mm); and (iv) mean daily temperature  $\geq 15.6$ °C (29).

*Cougarblight* also uses weather and phenology, but only to predict blossom infection by *E. amylovora*. *Cougarblight* was developed by Smith in Washington State because it was found that other models performed poorly in the Pacific Northwest (25). The inaccuracies of other forecasters were attributed to the estimation of daily mean temperatures and to the fact that inoculum pressure, which could vary between orchards because of disease history, was not considered (25,29). Inoculum pressure in this context is an estimate of overwintering inoculum and does not have a precise measurement of disease per unit area. In *Cougarblight*, the fire blight history of an orchard is explicitly used as an indicator of inoculum pressure (26). The categories of orchard fire blight history (potential for pathogen presence) used in *Cougarblight* Celsius version are given in Figure 1. Within each inoculum pressure category, risk levels of “low,” “caution,” “high,” and “extreme” are assigned based on a 4-day degree-hour total (base

Corresponding author: M. M. Dewdney; E-mail address: mmd38@cornell.edu

doi:10.1094/PHYTO-97-9-1164

This article is in the public domain and not copyrightable. It may be freely reprinted with customary crediting of the source. The American Phytopathological Society, 2007.

15.5°C). Blossom wetting events, including dew, are considered “triggers” for infection. Most growers and consultants consider management action when the risk level reaches high or extreme (25–27).

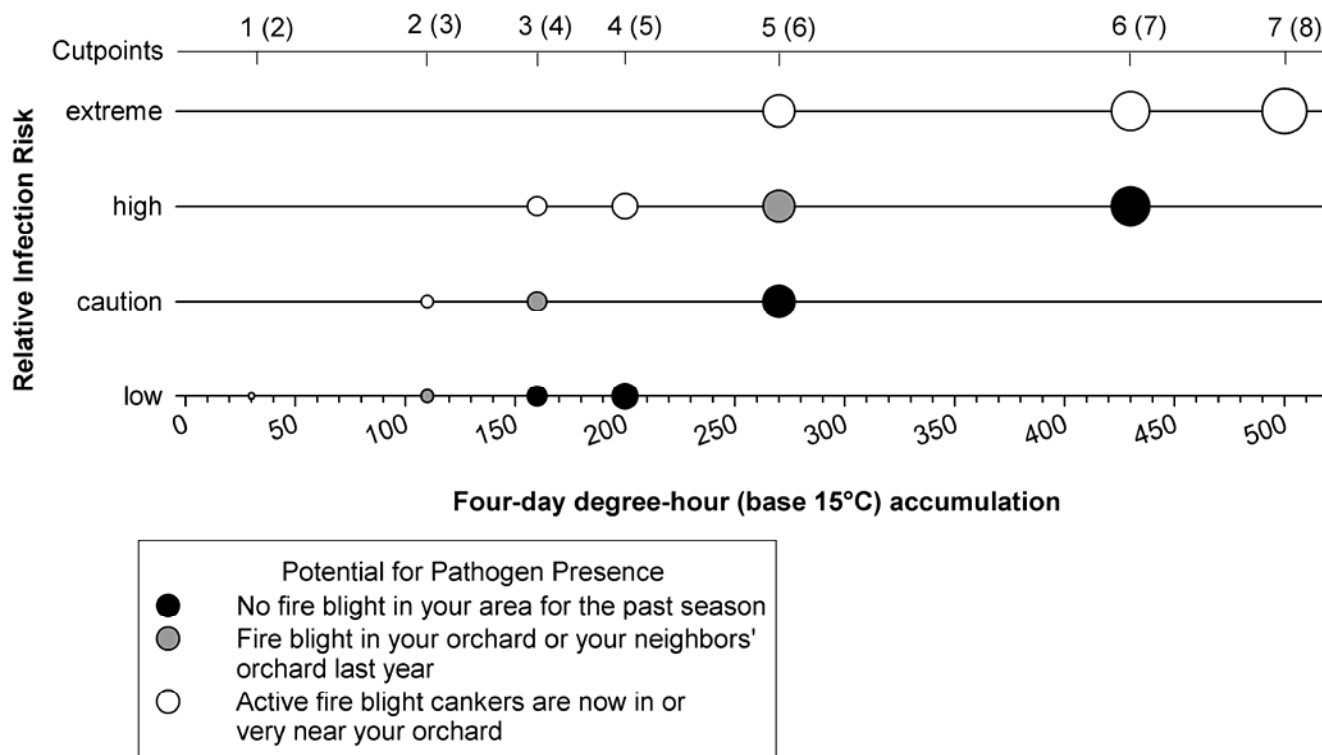
Despite the significant improvement to disease management brought about by fire blight forecasters, none, including *MARYBLYT* and *Cougarblight*, are perfect (3,8). The models tend to overpredict infection events, especially in less susceptible cultivars and in areas with infrequent fire blight events, resulting in needless applications of antibiotics (12,27,35). Our initial analysis indicated that *MARYBLYT* correctly predicted blossom blight 59 to 74% of the time depending upon geographic region, although the reasons for the regional differences were not entirely clear (7). In New York, the predictions by *MARYBLYT* and *Cougarblight* were highly correlated, indicating that there were only a few differences between the forecasters (4). Similarly, a comparison of several fire blight forecasters in Hungary, including *MARYBLYT* and *Cougarblight*, found good correlation of predictions between all the forecasters; however, the authors concluded that the greater number of predicted infection periods given by *Cougarblight* was a positive attribute (5). A correlation comparison, however, gives little indication of the accuracy of either forecaster, only that their predictions are similar. Another comparison of fire blight forecaster performance found that neither *MARYBLYT* nor *Cougarblight* performed well in Israel but that the two forecasters did not differ greatly in the number or timing of predicted infection periods. Conclusions were based on the coefficient of variation of the number of forecasted infections and the rate of correct infection periods (24). None of these approaches gives an idea of the number of times there were correct negative predictions. Another, intuitive, measure of forecaster accuracy would be to calculate the overall proportion of correct decisions. However, this type of accuracy calculation is dependent on the prevalence of disease in a data set, as discussed by Metz (16).

Receiver operating characteristic (ROC) curve analysis provides a more comprehensive method to evaluate and compare risk algorithms or forecasters and has been applied recently to plant diseases (11,32,33). ROC curve analysis has been widely used in the medical literature as a means to evaluate the performance of diagnostic tests in both clinical chemistry and radiology (9,16, 20,22). An ROC curve is informative for assessing how often a risk algorithm gives correct predictions, both positive and negative. Unlike the above comparisons used, in part, to determine forecaster accuracy, the ROC curve analysis does not depend on the proportions of cases and controls in a data set (16). A graphical method to evaluate and discriminate between diagnostic tests or modifications of the same test is provided. In choosing between two different forecasters or risk algorithms, the area under the ROC curve (AUC) can be used as a selective value. This is particularly useful when one is evaluating the predictive power of two forecasters, as in this study. The AUC measures the probability that, in randomly paired orchards, one with fire blight and the other without, the forecaster will correctly classify them (16,36).

The objectives of this study were to compare the predictive capacity of *MARYBLYT* and *Cougarblight* using ROC curve analysis, as well as to do a preliminary evaluation of the cutpoints (a group of thresholds) to determine which was optimal for accurately predicting an infection event.

### MATERIALS AND METHODS

In all, 243 historical data sets were collected from cooperators in British Columbia, England, Michigan, New York, Québec, Vermont, Washington state, and West Virginia. The data sets spanned the years from 1976 to 2002. Each data set contained daily maximum and minimum temperatures, precipitation, general apple phenology, and blossom blight incidence. Information in



**Fig. 1.** Cutpoints for *Cougarblight* and *Cougarblight and rain* are based on 4-day degree-hour (base 15°C) accumulations, levels of potential for pathogen presence, and relative blossom blight infection risk. The circles represent *Cougarblight* cutpoints and are labeled above the plot. Circle size increases with greater probability of disease. Cutpoint values in parentheses are for the *Cougarblight and rain* model which includes a precipitation threshold of  $\geq 0.25$  mm of rain or dew. Cutpoint 1 is for data sets that do not have precipitation as a trigger. In instances where levels of pathogen presence are at the same relative infection risk, only the circle for the most conservative fire blight inoculum pressure is shown.

the data sets varied in scale from a single orchard block to larger geographical areas of several hundred hectares. Some data sets had cultivar information that enabled more precise estimation of the phenology. For the analysis, the data sets first were used as a whole and then were grouped into large geographic regions and cultivar susceptibility levels to investigate whether these two variables had an effect on forecaster performance. Data sets from eastern North America were from Michigan, New York, Québec, Vermont, and West Virginia, and those from the west coast data sets were from British Columbia and Washington state. Data sets from England were analyzed as a group. The cvs. Elstar, Empire, Golden Delicious, Golden Russet, Jersey Mac, Jonafree, McIntosh, Mutsu, Spartan, and Vista Bella were defined as moderately susceptible (34). 'Red Delicious' was grouped with the moderately susceptible cultivars because there were not enough data sets to analyze it separately as a low-susceptibility cultivar. Very susceptible cultivars appearing in this study were Braeburn, Gingergold, Greening Rome, Idared, Jonagold, Jonathan, Lobo, Lodi, Paulared, Rome Beauty, Royal Gala, Yellow Transparent, and York (34). To ensure that disease was absent because conditions were unfavorable rather than as a result of a control measure, such as an antibiotic application, any data set with known antibiotic use and no observed symptoms was discarded before the ROC analysis. Following the conventions set in the medical literature, those data sets where fire blight was observed (*D+*) were classified as "cases" and those where fire blight was not observed were classified as "controls" (*D-*) (9,11,16). The simple presence or absence of fire blight in an orchard was used rather than a severity estimate because fire blight is considered such a potentially harmful disease that any "strikes" can be economically damaging (28,29). The prevalence of cases [P(*D+*)] for each data set group is presented in Table 1.

For each data set, the weather and phenology data for each year, location, and cultivar were entered into *MARYBLYT* and *Cougarblight*. A Microsoft Excel-based version of the *MARYBLYT* blossom blight module was developed to facilitate manipulations to the data sets during this study. The EIP and the other thresholds were calculated as described by Steiner and Lightner (29). The thresholds then were summed into the infection risk categories as in *MARYBLYT* (29) (Table 2). Similarly, Microsoft Excel was used to calculate the *Cougarblight* risk levels based on the risk tables presented in the 2002 Celsius version (26) (Fig. 1). When a dew event was entered into *MARYBLYT* and *Cougarblight*, 0.25 mm was automatically entered into the rain column,

TABLE 2. Cutpoints for the *MARYBLYT* model based on risk levels determined by the minimum values of thresholds as given by Steiner and Lightner (29)

Cutpoints	Risk levels <sup>b</sup>	Minimum levels reached for each threshold <sup>a</sup>		
		Wetness <sup>c</sup>	Temperature <sup>d</sup>	EIP <sup>e</sup>
1	M	+	...	...
2	M	...	+	...
3	M	...	...	+
4	H	...	+	+
5	H	+	...	+
6	H	+	+	...
7	I	+	+	+

<sup>a</sup> Blossoms are always considered open in this study and are not included in the cutpoint calculations; + indicates threshold reached.

<sup>b</sup> Risk levels do not necessarily progress directly from one to the next. Level of risk: M = medium, H = high, and I = infection imminent.

<sup>c</sup> Heavy dew, ≥0.25 mm of rain, or ≥2.5 mm of rain the day before.

<sup>d</sup> Mean daily temperature ≥15.6°C.

<sup>e</sup> Epiphytic infection potential (EIP) ≥100%.

TABLE 1. Prevalence of cases, area under the receiver operating characteristic curves (AUC), and comparison to the line of no discrimination for each forecaster and data set

Data set, prediction system	Sample size	Prevalence of cases	Point estimate of AUC <sup>a</sup>	95% Confidence interval of AUC		Significance <sup>c</sup>
				Asymptotic method <sup>a</sup>	Bootstrap method <sup>b</sup>	
All data						
<i>Cougarblight</i>	243	0.43	0.695	0.631–0.759	0.628–0.756	0.0000
<i>Cougarblight and rain</i>	...	...	0.668	0.597–0.738	0.597–0.737	0.0000
<i>MARYBLYT</i>	...	...	0.678	0.618–0.738	0.615–0.733	0.0000
Eastern North America						
<i>Cougarblight</i>	161	0.39	0.657	0.574–0.742	0.570–0.737	0.0005
<i>Cougarblight and rain</i>	...	...	0.630	0.536–0.725	0.531–0.721	0.0045
<i>MARYBLYT</i>	...	...	0.681	0.610–0.752	0.606–0.747	0.0000
England						
<i>Cougarblight</i>	34	0.56	0.861	0.741–0.981	0.696–0.950	0.0002
<i>Cougarblight and rain</i>	...	...	0.858	0.730–0.986	0.676–0.953	0.0003
<i>MARYBLYT</i>	...	...	0.740	0.576–0.905	0.547–0.879	0.0097
West coast						
<i>Cougarblight</i>	48	0.48	0.755	0.623–0.887	0.601–0.865	0.0016
<i>Cougarblight and rain</i>	...	...	0.672	0.515–0.829	0.503–0.815	0.035
<i>MARYBLYT</i>	...	...	0.623	0.461–0.772	0.449–0.763	0.12
Moderately susceptible cultivars						
<i>Cougarblight</i>	87	0.30	0.761	0.653–0.869	0.639–0.854	0.0001
<i>Cougarblight and rain</i>	...	...	0.706	0.564–0.848	0.543–0.835	0.0020
<i>MARYBLYT</i>	...	...	0.698	0.601–0.795	0.597–0.787	0.0017
Very susceptible cultivars						
<i>Cougarblight</i>	84	0.58	0.621	0.503–0.740	0.499–0.733	0.052
<i>Cougarblight and rain</i>	...	...	0.567	0.446–0.687	0.444–0.681	0.29
<i>MARYBLYT</i>	...	...	0.688	0.581–0.794	0.575–0.789	0.0010

<sup>a</sup> Based on method of DeLong et al. (6).

<sup>b</sup> Based on 20,000 iterations of the bias-corrected accelerated bootstrap method (36).

<sup>c</sup> Statistical significance of difference from 0.5 (*P* value); *z* value for the null *H*<sub>0</sub>: AUC = 0.5 or line of no discrimination based on the Mann-Whitney *U* statistic

$$(U) (36). \text{ The two-tailed test statistic is given by } z = \frac{\left| AUC - \frac{n_1(n_1 + n_2 + 1)}{2} \right| - 0.5}{SE_U} \text{ where } SE_U = \sqrt{\frac{n_1 n_2}{12} \left[ (n_1 + n_2 + 1) - \frac{\sum_i (t_i^3 - t_i)}{(n_1 + n_2)(n_1 + n_2 - 1)} \right]}, \text{ } t_i = \text{ the number of tied values in } i\text{th set of ties.}$$

as is done in the *MARYBLYT* program (29). The amount of blossom wetting needed for infection is unspecified for the *Cougarblight* model (25–27); therefore, the same minimum amount of precipitation as in *MARYBLYT* was used.

The output of the individual data sets from *MARYBLYT* and *Cougarblight* was arranged into a separate file for each of the forecasters. For days during bloom, when blossom infection could occur, the risk level was calculated for each day and the highest of the risk levels was used for the data set. Thus, there was a prediction risk and observed infection or disease value for each data set. Predictions of infection then could be made for each of the seven or eight cutpoints to construct the ROC curve. For example, if a data set had the highest *MARYBLYT* predicted risk (at cutpoint 7), then infection would be predicted for any cutpoints (defined below) from 1 to 7 because an observed cutpoint 7 indicates that all lower conditions for infection were surpassed. If a data set had the next highest predicted risk (cutpoint 6) as the highest risk value, then infection would be predicted for cutpoints 1 through 6 but not 7, because the risk was not great enough to require the highest risk for an infection prediction. Conversely, if the highest risk in a data set was cutpoint 1, then infection would be predicted only for cutpoint 1 as the threshold and not for cutpoints 2 through 7 because the more stringent conditions were not met for the data set. For the evaluation of *Cougarblight*, the same days from individual data sets as in *MARYBLYT* were used to compare the forecasters on an as equitable basis as possible. To confirm that a bias toward *MARYBLYT* performance was not introduced into the analysis, we took a random subset equaling 20% of the total number of data sets in proportion to geographic region. The day with the highest risk point, as determined with *Cougarblight*, was selected and used to compare forecasts with *MARYBLYT*. The subset was analyzed as described below. Cutpoints were calculated for the selected day from individual data sets. The results then were compiled into the different regional and cultivar susceptibility data sets. The cutpoints or forecaster outcomes then were matched with the disease incidence of each data set.

A cutpoint (*T*) is a predefined value that identifies the prediction of an infection event (36). Cutpoints are considered to be real numbers, where the highest number has a greater likelihood of being associated with disease, but there is no need to progress directly from one cutpoint to the next. For each forecaster, a range of cutpoints was defined (Table 2; Fig. 1). Depending on the risk algorithm being tested, cutpoints can be either a collection of model thresholds, as in *MARYBLYT*, or single thresholds, as in *Cougarblight*. *MARYBLYT* cutpoints were based on how many thresholds had exceeded their minimum values and are defined in Table 2. The cutpoint order was based on calculations of sensitivity and specificity, originally done by hand, for our initial ROC curve analysis of *MARYBLYT* (7). The cutpoints were confirmed to be unique and in the same order by the program AccuROC (36). Because, in this study, blossoms were considered to be always open, the blossom threshold was not included in the *MARYBLYT* cutpoint definitions. Cutpoint definitions were not as simple with *Cougarblight*. Because no information about the potential pathogen presence criteria in *Cougarblight* was available for the majority of data sets, the cutpoints were constructed to evaluate all three levels of potential pathogen presence simultaneously (Fig. 1). For example, cutpoint 1 for *Cougarblight* without the rain threshold includes all levels of potential pathogen presence in the risk level “low” up to 30 degree-hours. Cutpoint 2 includes up to 110 degree-hours. This is equivalent to the potential pathogen presence levels “no fire blight in your area for the past season” and “fire blight in your orchard or your neighbor’s orchard last year” at risk level “low”, as well as “active fire blight cankers are now in or very near your orchard” at risk level caution. This is similar to how cutpoints are assigned in radiology (9). The remaining cutpoints were defined based on the degree-hour summations (Fig. 1) in the manner described above (27). In

addition, precipitation was indicated to be the trigger event in fire blight infection if the risk levels were high enough (25–27). To better evaluate the effect of the rain threshold (precipitation  $\geq 0.25$  mm) on forecaster reliability, the data sets were evaluated with and without the trigger event. The two model permutations were termed “*Cougarblight and rain*” and “*Cougarblight*,” respectively.

For each data set, the predicted infection based on each of the cutpoint thresholds was compared with the observation of disease presence (“actual disease”) and the results categorized into four decision outcomes: true positive, true negative, false positive, and false negative (16,17). From the four decision outcomes, category proportions were calculated. For example, the true positive proportion (TPP), also known as the sensitivity of a test, is the number of correctly classified cases over the total number of cases and can be defined by the conditional probability of  $P(T+|D+)$  (16). Similarly, the false positive proportion (FPP) is the number of controls incorrectly predicted to be diseased over the total number of controls, which is defined by the conditional probability  $[P(T+|D-)]$ . The FPP is denoted as  $1 - \text{specificity}$ , where the specificity is the true negative proportion (TNP) or  $[P(T-|D-)]$  (16). The sensitivity versus  $1 - \text{specificity}$  of all the possible cutpoints of both fire blight forecasters were plotted as ROC curves. In addition to calculating the sensitivity and  $1 - \text{specificity}$  for each cutpoint, the frequency of cases and controls at each cutpoint of *MARYBLYT* and *Cougarblight* were plotted (Figs. 2 to 4).

To interpret the results of an ROC curve analysis, it is necessary to know certain fundamental features of the analysis. By definition, an ROC curve passes through the point 0,0, where all orchards are predicted to be negative for fire blight at the highest cutpoint, and the point 1,1, meaning that all orchards are predicted positive for fire blight at the lowest cutpoint (11,16,17). As the ROC curve approaches the point 0,1, the performance of a forecaster improves and exhibits both beneficial sensitivity and specificity characteristics (11,16,36). To compare blossom blight forecasting ability between *MARYBLYT* and *Cougarblight* in a more quantifiable manner, the AUC was used as a discriminating value. Bamber (1), with elaboration from Hanley and McNeil (9), recognized that this probability is equal to the nonparametric Mann-Whitney *U* statistic (*U*) from which standard errors can be calculated. Accordingly, the closer an AUC is to 1.0, the better the forecaster is at correctly classifying the case and control. The Mann-Whitney *U* statistic approaches normality with increased sample size; however, the actual probability distribution at small sample size of the AUC is unknown and generally is asymmetric at the extremes near 0 and 1. This can result in upper bounds  $>1$  with large standard errors. A way to more accurately estimate confidence intervals is to use bootstraps, the best being the bias-corrected accelerated method (6,19,36). We used the program AccuROC (36) to calculate the sensitivity and  $1 - \text{specificity}$ , the AUCs, and the 95% confidence intervals (CIs), both asymptotically and via bootstraps, for each curve.

Another feature of ROC analysis is the “line of no discrimination,” which passes through the points 0,0 and 1,1 and has an AUC that is equal to the probability of 0.5 or 50%. If an ROC curve falls along the line of no discrimination, the forecaster does not predict an infection period any better than chance (11). We used two methods to determine whether a forecaster was performing significantly better than chance in a specific test. The first test was to observe whether the 95% CIs, either asymptotic or bootstrap, excluded the value of 0.5. The second was to test the null hypothesis that the AUC was not significantly different from 0.5 based on a *z* statistic and two-tailed *P* value derived from the Mann-Whitney *U* statistic (11,36).

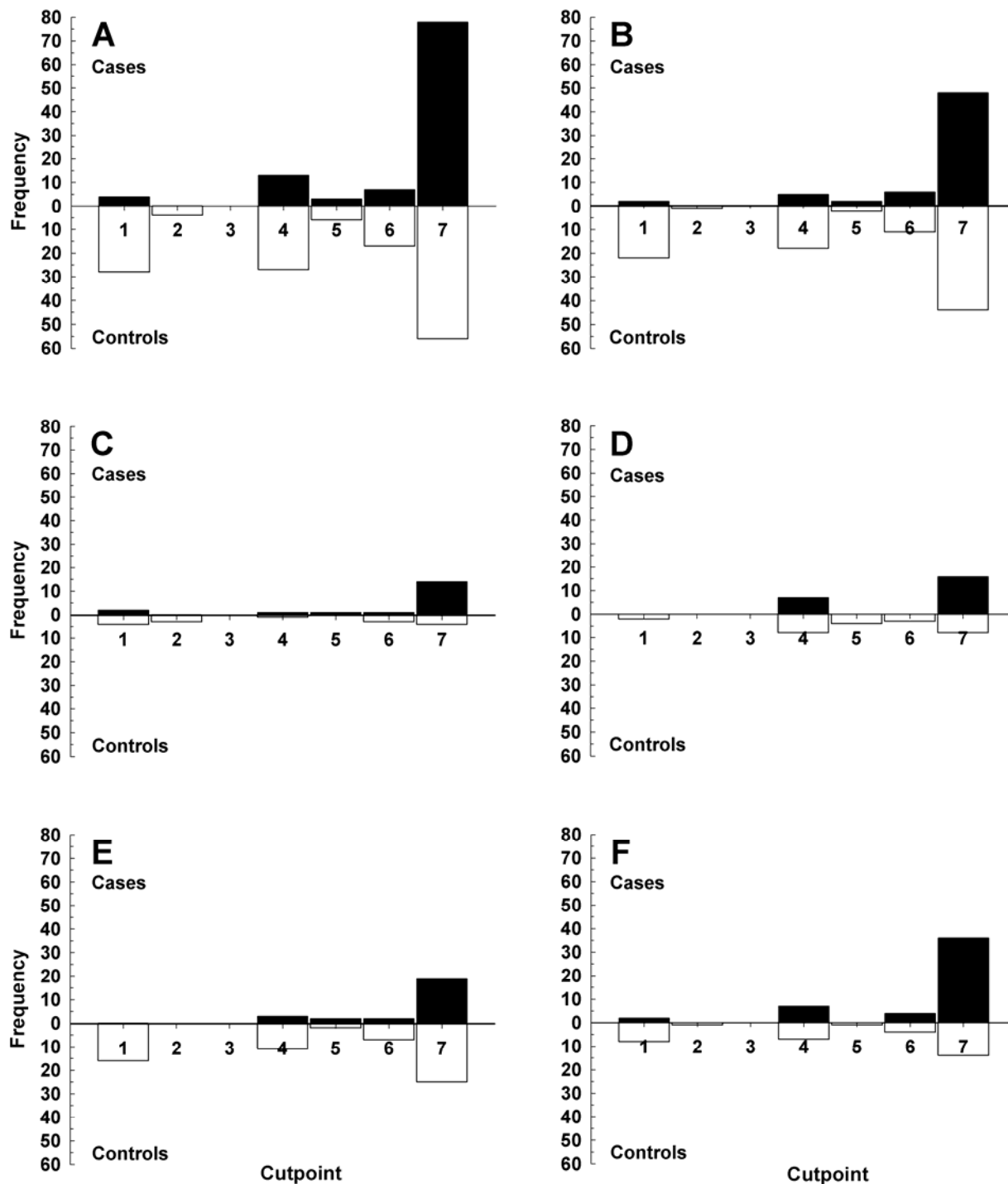
The performance of *MARYBLYT* and *Cougarblight* forecasts was compared for the same geographic regions and cultivar susceptibility. The first test compared forecaster performance by

whether the 95% CIs of the AUCs overlapped for curves generated from the same data sets. A more formalized test constructed contrasts for the three curves, that allowed for the calculation of  $\chi^2$  values and *P* values to assess whether the contrast was significant (6,36). Further pairwise contrasts were constructed to ensure that there were no differences between individual curves. In addition, correlations between individual curves were calculated with the AUC and the variance-covariance matrix of each curve based on the Mann-Whitney *U* statistic (*U*) (6,36). A final question of interest was which forecaster cutpoint was most accurate. When false positives and false negatives are considered to be of equal importance, the cutpoint that is closest to the point

(0,1) is considered the best (10,16,37). Youden's index ( $J = \text{sensitivity} + \text{specificity} - 1$  or  $J = \text{TPP} - \text{FPP}$ ), a measure of the overall non-error rate for each cutpoint, was used to determine the optimal cutpoint in conjunction with the ROC curve analysis. The index has a range of 0 to 1, which is equivalent to no discriminatory power to perfect discriminatory power, respectively (10,37).

## RESULTS

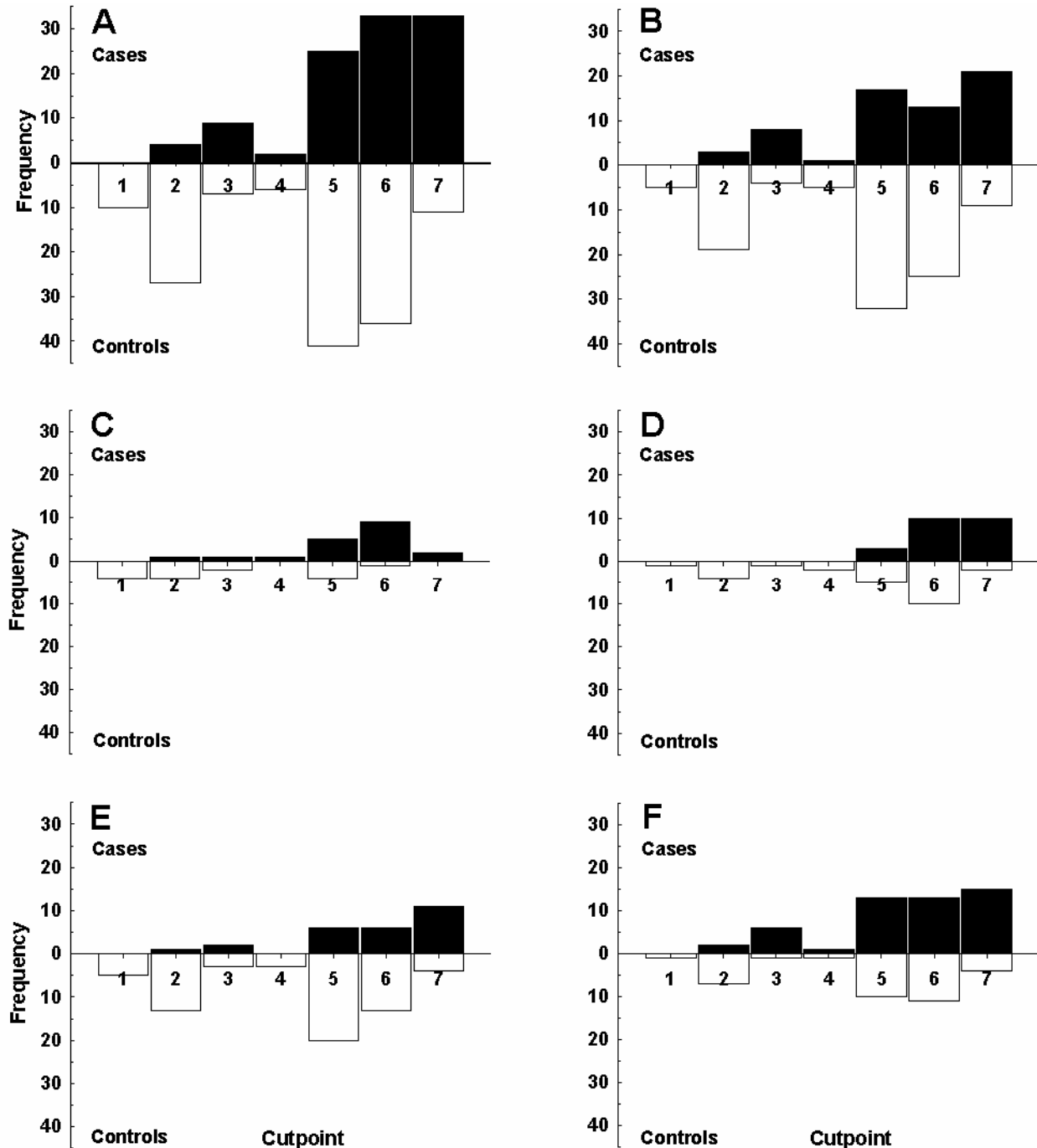
Based on ROC curve analysis, *MARYBLYT* and *Cougarblight* predicted infection risks equally accurately overall. To begin the analysis, the frequency distributions of cases and controls for



**Fig. 2.** Frequency distributions of cases and controls classified by *MARYBLYT* cutpoints for each data set. **A**, All data (243 points); **B**, eastern North America (161 points); **C**, England (34 points); **D**, west coast of North America (48 points); **E**, moderately susceptible cultivars (87 points); and **F**, very susceptible cultivars (84 points).

*MARYBLYT*, *Cougarblight*, and *Cougarblight and rain* were plotted as functions of their cutpoints in Figures 2, 3, and 4, respectively. Frequency distributions are presented for all regional and cultivar susceptibility data sets separately. In general, for both models, the distributions between the cases and the controls were overlapping. Ideally, the distributions would not overlap, with the cutpoint separating the two distributions being selected as optimal for forecast accuracy. In *MARYBLYT*, the EIP alone (cutpoint 3) never occurred for either cases or controls; therefore, it was not included in further ROC analysis. In the data sets from the west coast of North America (Fig. 2D) and moderately susceptible cultivars (Fig. 2E), neither cases nor controls were categorized by cutpoint 2 of *MARYBLYT*. The majority of the cases were classified by cutpoint 7 and the controls were spread among cutpoints

for each data set. For the two permutations of *Cougarblight*, without and with the rain threshold (Figs. 3 and 4), both the cases and controls were spread among the cutpoints for each data set. The distributions were similar for the two model permutations because the cutpoints for *Cougarblight* are equivalent to cutpoints 2 to 8 in *Cougarblight and rain*. Having the rain threshold separated in such a manner, however, did provide a unique opportunity to see where precipitation can play an important role for fire blight prediction. The majority of cases that were classified by cutpoint 1 (no rain event as a trigger of infection) in *Cougarblight and rain* were from cutpoints 5 and 6 of *Cougarblight*. No cases from cutpoints 1 through 4 were reclassified using the rain threshold and only a few were reclassified from cutpoint 7 of *Cougarblight* across region or cultivar susceptibility. The greatest



**Fig. 3.** Frequency distributions of cases and controls classified by *Cougarblight* cutpoints for each data set. **A**, All data (243 points); **B**, eastern North America (161 points); **C**, England (34 points); **D**, west coast of North America (48 points); **E**, moderately susceptible cultivars (87 points); and **F**, very susceptible cultivars (84 points).

numbers of reclassifications were found in the eastern and very susceptible cultivar data sets (Figs. 3B and F and 4B and F). In contrast, controls classified by cutpoint 1 in *Cougarblight and rain* were from all *Cougarblight and rain* cutpoints although, once again, the majority were from cutpoints 5 to 7. Many of the reclassified controls were from the eastern data set when grouped by region.

Equivalent numbers of controls were reclassified upon comparison of moderately susceptible cultivars to very susceptible cultivars, although the distributions were different. Controls from all *Cougarblight and rain* cutpoints were reclassified in the moderately susceptible cultivar data set, whereas all of the reclassified controls of the very susceptible cultivars were from cutpoints 5 to 7 of

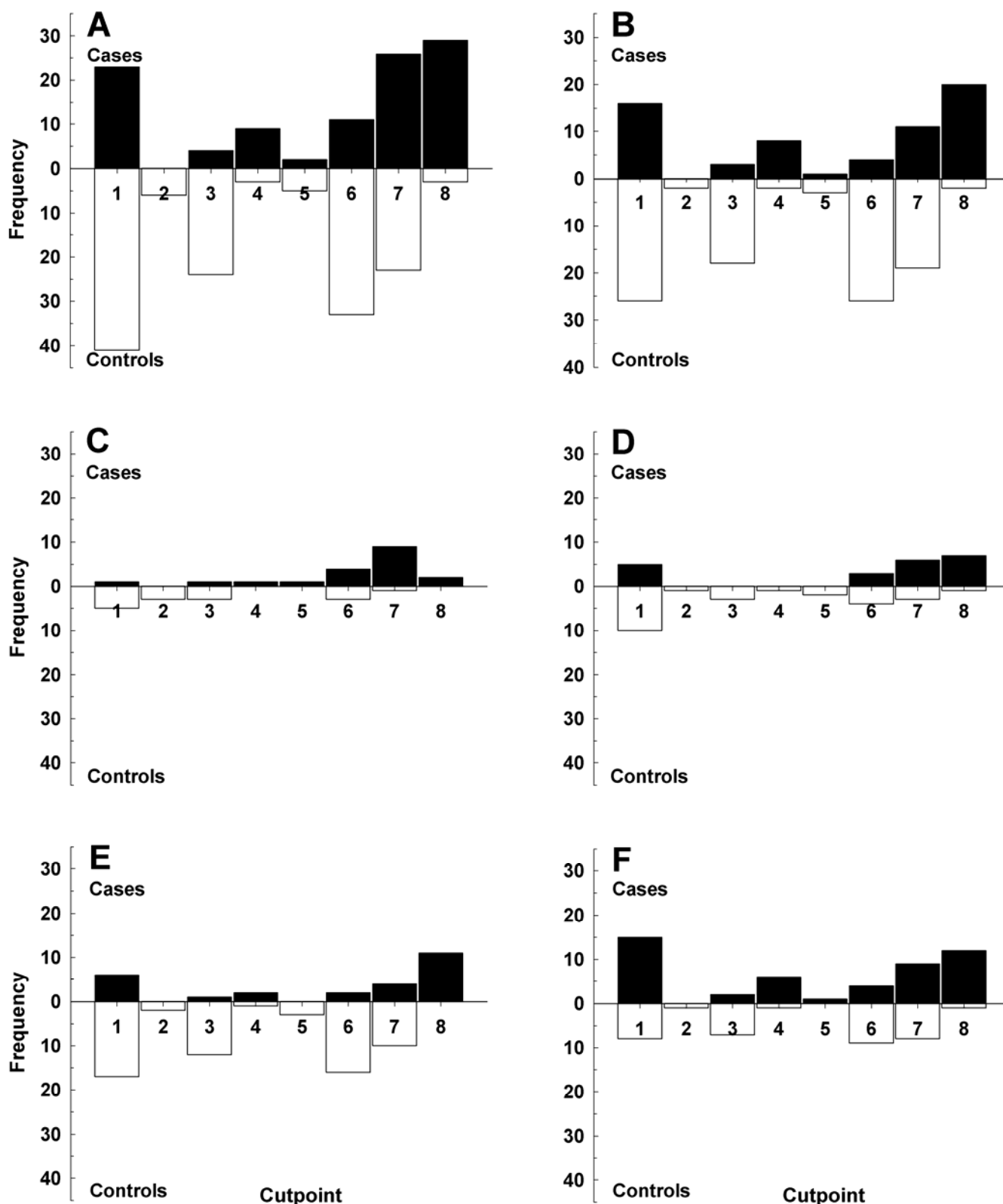
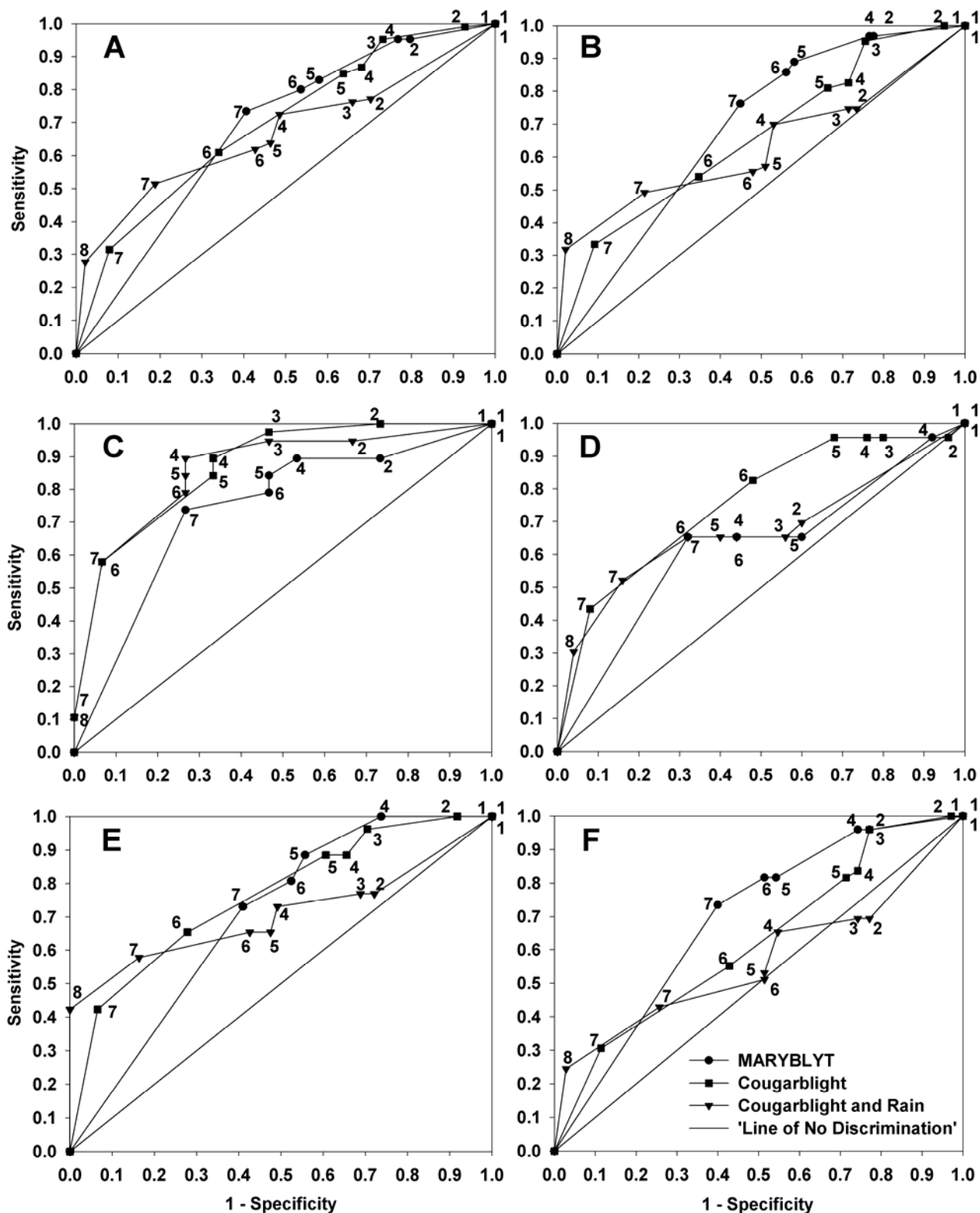


Fig. 4. Frequency distributions of cases and controls classified by *Cougarblight and rain* cutpoints for each data set. A, All data (243 points); B, eastern North America (161 points); C, England (34 points); D, west coast of North America (48 points); E, moderately susceptible cultivars (87 points); and F, very susceptible cultivars (84 points).

*Cougarblight*. Very few cases or controls were reclassified by the rain threshold in *Cougarblight* and rain in the England data set.

When the ROC curves with all data sets were plotted (Fig. 5A) and the AUCs calculated (Table 1), both *Cougarblight* and *MARYBLYT* were shown to predict fire blight infection significantly better than chance ( $P = 0.05$ ). There were no significant differences in predictive ability among the forecasts of the two

forms of *Cougarblight* and *MARYBLYT* according to the three-way contrast, which was confirmed by the overlapping 95% CIs from each curve (Tables 1 and 3). When the AUCs of *Cougarblight* and *MARYBLYT* forecasts were compared for the data subset where *Cougarblight* was used to determine the day with the highest risk point, no significant differences in predictive ability were found (data not shown). When the model forecasts were



**Fig. 5.** Receiver operating characteristic plots with curves for *MARYBLYT*, *Cougarblight*, and *Cougarblight* by data set. A, All data; B, eastern North America; C, England; D, west coast of North America; E, moderately susceptible cultivars; and F, very susceptible cultivars. The cutpoints for the models are labeled on the lines.

compared using pairwise contrasts, no further significant differences were found. Correlations among *Cougarblight*, *Cougarblight and rain*, and *MARYBLYT* forecasts (Table 4) showed that the two forms of *Cougarblight* forecasts only had a correlation of 0.47 despite the small difference between cutpoints. The highest correlation (0.86) for all data sets was between *Cougarblight and rain* and *MARYBLYT* forecasts with the west coast data set.

In all of the regional data sets, both *Cougarblight* and *MARYBLYT* forecast blossom blight infection periods significantly better than chance ( $P < 0.05$ ), with the exception of *MARYBLYT* when applied to the data set from the west coast of North America ( $P = 0.12$ ) (Fig. 5B to D). Also, the 95% CIs of each ROC curve AUC (Table 1) did not contain 0.5 within their limits, with the exception of *MARYBLYT* forecasts with the west coast data set. Similarly, when using the west coast data set, the ROC curve of *Cougarblight and rain* forecasts had a 95% CI that almost encompassed 0.5 and had a  $P$  value that approached 0.05 (Table 1). The ROC curves did not differ ( $P > 0.05$ ) among the forecasters according to three-way or pairwise contrasts among the east and west coasts of North America (Table 3). However, the  $\chi^2$  value of the three-way contrast from the England data set was significant at the 10% significance level (90% confidence). In addition, both forms of the *Cougarblight* model predicted blossom blight significantly better ( $P > 0.05$ ) than *MARYBLYT* in the England data set (Table 3). The AUCs of the two forms of *Cougarblight* were

similar in shape and were statistically equivalent based on the  $\chi^2$  test (Fig. 5C). In Figure 5, the ROC curve for *Cougarblight and rain* crossed the ROC curves of the other two models in every data set, and the only time that the *MARYBLYT* ROC curve was not crossed was in the England data set. Correlations between the forecasts for each geographic region data set are presented in Table 4. Overall, the highest correlations of forecasts were from the England data set, ranging between 0.7 and 0.8. The lowest correlations of predictions were found using the west coast data set between the two forms of *Cougarblight* and between *Cougarblight* and *MARYBLYT*. Within range of the same data set, the single highest correlation was found between *Cougarblight and rain* and *MARYBLYT*. Correlations between the forecaster's predictions for the eastern data set were fairly similar and ranged between 0.45 and 0.60.

Cultivar susceptibility adjustments are not features of either forecaster; however, there were concerns that it might affect forecaster performance. ROC curves for both moderately and very susceptible cultivars showed that *MARYBLYT* predicted significantly better ( $P < 0.05$ ) than chance because 0.5 was not contained in the 95% CI of the AUCs (Table 1; Fig. 5E and F). The situation for *Cougarblight* differed in that, with the very susceptible cultivar data set, neither form of *Cougarblight* had an AUC that was significantly  $>0.5$ . The 95% CIs of both AUCs contained 0.5 (Table 1), although *Cougarblight* without the rain threshold

TABLE 3. The  $\chi^2$  values of three-way and pairwise contrasts among the areas under the receiver operating characteristic curves (AUC) of *MARYBLYT*, *Cougarblight*, and *Cougarblight and rain* for geographic regions and cultivar susceptibilities

Data set	Three-way comparison <sup>a</sup> ( $P$ value)	<i>MARYBLYT</i> versus <i>Cougarblight</i> <sup>b</sup> ( $P$ value)	<i>MARYBLYT</i> versus <i>Cougarblight and rain</i> ( $P$ value)	<i>Cougarblight</i> versus <i>Cougarblight and rain</i> ( $P$ value)
All data	0.585 (0.746)	0.227 (0.634)	0.120 (0.729)	0.585 (0.444)
Eastern North America	1.244 (0.537)	0.334 (0.563)	1.243 (0.265)	0.400 (0.527)
England	5.126 (0.077)	4.618 (0.032)	4.140 (0.042)	0.0066 (0.935)
West Coast	3.186 (0.203)	2.100 (0.147)	1.351 (0.245)	0.720 (0.396)
Moderately susceptible cultivars	1.320 (0.517)	1.094 (0.296)	0.0169 (0.897)	0.807 (0.369)
Very susceptible cultivars	4.948 (0.084)	1.026 (0.311)	4.899 (0.027)	0.646 (0.422)

<sup>a</sup> Based on method of DeLong et al. (6) and calculated using AccuROC (36) with two degrees of freedom.

<sup>b</sup> One degree of freedom for all two-way contrasts.

TABLE 4. Correlation coefficients of the comparisons among the forecasts of *MARYBLYT*, *Cougarblight*, and *Cougarblight and rain* for each geographic region and cultivar susceptibility<sup>a</sup>

Data set, prediction system	<i>Cougarblight</i>		<i>Cougarblight and rain</i>		<i>MARYBLYT</i>	
	$r$	$P$ value	$r$	$P$ value	$r$	$P$ value
All data						
<i>Cougarblight</i>	1.00	<0.0001	...	...	...	...
<i>Cougarblight and rain</i>	0.46	<0.0001	1.00	<0.0001	...	...
<i>MARYBLYT</i>	0.40	<0.0001	0.57	<0.0001	1.00	<0.0001
Eastern North America						
<i>Cougarblight</i>	1.00	<0.0001	...	...	...	...
<i>Cougarblight and rain</i>	0.58	<0.0001	1.00	<0.0001	...	...
<i>MARYBLYT</i>	0.45	<0.0001	0.45	<0.0001	1.00	<0.0001
England						
<i>Cougarblight</i>	1.00	<0.0001	...	...	...	...
<i>Cougarblight and rain</i>	0.77	<0.0001	1.00	<0.0001	...	...
<i>MARYBLYT</i>	0.74	<0.0001	0.73	<0.0001	1.00	<0.0001
West coast						
<i>Cougarblight</i>	1.00	<0.0001	...	...	...	...
<i>Cougarblight and rain</i>	0.14	0.171	1.00	<0.0001	...	...
<i>MARYBLYT</i>	0.22	0.066	0.86	<0.0001	1.00	<0.0001
Moderately susceptible cultivars						
<i>Cougarblight</i>	1.00	<0.0001	...	...	...	...
<i>Cougarblight and rain</i>	0.58	<0.0001	1.00	<0.0001	...	...
<i>MARYBLYT</i>	0.35	0.0005	0.52	<0.0001	1.00	<0.0001
Very susceptible cultivars						
<i>Cougarblight</i>	1.00	<0.0001	...	...	...	...
<i>Cougarblight and rain</i>	0.37	0.0003	1.00	<0.0001	...	...
<i>MARYBLYT</i>	0.35	0.0006	0.56	<0.0001	1.00	<0.0001

<sup>a</sup> For each forecast system,  $r$  was calculated by AccuROC (36) and  $P$  values were calculated from the test statistic  $t = r\sqrt{n-2}/\sqrt{1-r^2}$ .

contained 0.5 only within the 95% CI calculated by bootstraps. The only ROC curve to fall below the line of discrimination in the entire study was *Cougarblight and rain* using the very susceptible cultivar data set (Fig. 5F). When the three AUCs were compared for the cultivar susceptibility data sets, there were no significant differences ( $P < 0.05$ ) among them (Table 3). Similarly, no pairwise differences were observed between the predictions of the three models for the moderately susceptible cultivar data set. When compared in a pairwise manner for the very susceptible cultivars data set, the AUCs of *Cougarblight and rain* and *MARYBLYT* forecasts were significantly different ( $P = 0.03$ ). For both cultivar susceptibility data sets, the correlation between *Cougarblight* and *MARYBLYT* was  $< 0.4$ . Correlations between the predictions of the two *Cougarblight* models were higher for the moderately susceptible cultivar data set than for the very susceptible cultivar data set (Table 4). The correlation between *MARYBLYT* and *Cougarblight and rain* forecasts for the very susceptible cultivar data set was unexpectedly higher than that for the moderately susceptible cultivar data set because there were significant differences between the AUCs of the ROC curves in the one data set (Tables 3 and 4).

The optimal cutpoint for *MARYBLYT*, as determined by Youden's index, was cutpoint 7 in all but one data set; the data set for moderately susceptible cultivars had an optimal cutpoint of 5 (Table 5). Cutpoint 7 (Table 2) was equivalent to imminent infection in the *MARYBLYT* program, and all three thresholds had exceeded their minimum values. The highest value ( $J = 0.47$ ) for Youden's index in *MARYBLYT* was found with the England data set. In both versions of *Cougarblight*, the optimal cutpoint varied among data sets with no consistent pattern (Fig. 1; Table 5). The cutpoint with the highest Youden's index among the models was most often found in *Cougarblight and rain*.

## DISCUSSION

A useful plant disease forecaster needs to reliably discriminate whether an infection event will occur or not. Our study has shown that both *MARYBLYT* and *Cougarblight* can be used to distinguish whether a fire blight infection event will occur, but there clearly is room for considerable improvement. In most situations, there were no significant differences in the performance, as measured by the AUC, of *MARYBLYT* versus *Cougarblight* forecasts. There was variability among geographic regions and cultivar susceptibilities, but no single forecaster outperformed the others based on three-way contrasts of AUCs. Only two situations, the England and the very susceptible cultivar data sets, had significant differences in performance between the two forecasters' predictions when pairwise comparisons were used (Tables 2 and 3). The best overall performance of *Cougarblight* occurred with the England data set and the best performance of *MARYBLYT* was in the eastern data set. *MARYBLYT* and *Cougarblight and rain* did not perform well with the west coast data set. Neither version of *Cougarblight* did well with very susceptible cultivars; therefore,

caution should be exercised when relying on it exclusively in situations where many valuable, very susceptible cultivars are at risk. In contrast, *MARYBLYT* performed equally well with both cultivar susceptibility data sets. The variability in model performance among regions is still unexplained but possibly could reflect differences in environment, population variation of *E. amylovora*, or cultivar prevalence in a region, as is discussed further. Some of the variability is due to the disparity among data sets that were collected, where some were specific to block and cultivar whereas others encompassed a geographic region. However, not all of the variability is explained by this factor. A major consequence of the variability in the data sets is that the standard errors of the AUCs tended to be large, possibly masking differences in forecaster accuracy between regions that would be significant otherwise (data not shown).

When the frequency of cases and controls were plotted against the cutpoints of a forecaster typically the two distributions overlap. Ideally, the majority of controls would be classified by lower cutpoints and cases by higher ones, with the intersection of the distributions being considered the optimal cutpoint (11). The difficulty with our data is that, although the majority of cases were ranked by upper cutpoints, the controls were fairly evenly distributed among cutpoints of both forecasters, with the exception of the highest cutpoints of the two versions of *Cougarblight*. When looking more specifically at *MARYBLYT*, cutpoint 3 (EIP alone) never classified either cases or controls in any data set (Fig. 2). It is likely that this is in part because EIP is highly correlated with temperature and, thus, perhaps not a unique variable in cool temperatures (29). A full day of temperatures  $> 18^{\circ}\text{C}$  is needed to accumulate enough degree-hours to cross the EIP threshold. EIP is an estimate of *E. amylovora* population increase and spread (29). A temperature effect on bacteria populations in the field is likely to take several hours; therefore, a substantial population change may not occur on the day of EIP accumulation. This may mean that EIP calculated from the current day's heat accumulation does not give the best forecasts. In addition, cutpoint 2, mean daily temperature, did not classify any cases or controls in either the west coast or moderately susceptible cultivar data sets, suggesting that mean daily temperature alone also was a relatively poor infection event predictor. Clearly, both daily mean temperature and EIP contributed to prediction accuracy, because when one was removed from the prediction criteria (for example, changing cutpoint 7 to cutpoints 6 or 5), the frequency at which cases were classified fell precipitously. The cutpoint that classified the second greatest number of cases was number 4, daily mean temperature and EIP, and was the only cutpoint other than number 7 that classified any cases with the west coast data set. Even though the sensitivity of cutpoint 4 generally was higher than cutpoint 7, the specificity was much lower (Fig. 5), showing that, whereas temperature variables were an important factor in determining a correct prediction of infection, they were not as effective in correctly classifying controls. The effects of the tem-

TABLE 5. Optimal cutpoints (Cut) for *MARYBLYT*, *Cougarblight*, and *Cougarblight and rain* based on the highest Youden's index (J) value in each data set

Data set	<i>MARYBLYT</i>			<i>Cougarblight</i>			<i>Cougarblight and rain</i>		
	Cut <sup>a</sup>	TPP, FPP <sup>b</sup>	J (95% CI) <sup>c</sup>	Cut	TPP, FPP	J (95% CI)	Cut	TPP, FPP	J (95% CI)
All data	7	0.733, 0.406	0.328 (0.211, 0.444)	6	0.610, 0.341	0.269 (0.148, 0.390)	7	0.514, 0.188	0.326 (0.201, 0.451)
Eastern North America	7	0.762, 0.449	0.313 (0.172, 0.454)	7	0.333, 0.092	0.242 (0.059, 0.424)	8	0.318, 0.020	0.297 (0.155, 0.440)
England	7	0.737, 0.267	0.470 (0.173, 0.768)	4	0.895, 0.333	0.561 (0.287, 0.836)	4	0.895, 0.267	0.628 (0.370, 0.886)
West coast	7	0.652, 0.320	0.332 (0.065, 0.599)	7	0.435, 0.08	0.355 (0.092, 0.618)	7	0.522, 0.160	0.362 (0.093, 0.630)
Moderately susceptible cultivars	5	0.885, 0.557	0.327 (0.161, 0.494)	6	0.654, 0.279	0.375 (0.179, 0.571)	8	0.423, 0.000	0.423 (0.334, 0.513)
Very susceptible cultivars	7	0.735, 0.400	0.338 (0.129, 0.540)	7	0.306, 0.114	0.192 (0, 0.412)	8	0.245, 0.029	0.216 (0.031, 0.402)

<sup>a</sup> Best cutpoint for a data set is based on highest J. An explanation of the cutpoints is found in Table 1 and Figure 1.

<sup>b</sup> TPP = true positive proportion, FPP = false positive proportion.

<sup>c</sup>  $J = \text{TPP} - \text{FPP}$  (=sensitivity + specificity - 1). CI = confidence interval.  $95\% \text{ CI} = J \pm 1.96(\text{SE}_J)$  where  $\text{SE}_J = \sqrt{\frac{\text{TP} \times \text{FP}}{(\text{TP} + \text{FP})^3} + \frac{\text{FN} \times \text{TN}}{(\text{FN} + \text{TN})^3}}$ , where TP = true positives, FP = false positives, FN = false negatives, and TN = true negatives (37).

perature-based variables also were visible on the ROC curves regardless of data set (Fig. 5), as shown by the greater specificity but reduced sensitivity of cutpoint 7 compared with cutpoints 5 and 6.

Several studies have shown that a rain event is needed to move *E. amylovora* from the stigma surface where the bacteria multiply to the hypanthium (floral cup), the main site of blossom infection (21,30,31), leading to the conclusion that rain could be a trigger event for an infection. When the effects of the *Cougarblight* rain threshold were investigated, it was apparent that there was an association between the 4-day degree-hour accumulations and precipitation. No cases from cutpoints 1 to 4, those with degree-hour accumulations <200, of *Cougarblight* (without rain) were predicted by cutpoint 1 (no precipitation event) of *Cougarblight and rain*. The majority of cases reclassified by cutpoint 1 were from cutpoints 5 and 6 of *Cougarblight* (Figs. 3 and 4). The effect on controls when the rain threshold was added to *Cougarblight* was more uniform; the majority of the redistributed controls not coming from a particular cutpoint. Reclassification of controls did not significantly improve the predictive ability of *Cougarblight*, even though many controls were more appropriately categorized because so many of the cases also were classified by cutpoint 1, balancing out the effect. The association of the rain threshold and the 4-day degree-hour accumulations also can be visualized with the ROC curves (Fig. 5). The ROC curves for *Cougarblight and rain* cross those for *Cougarblight* in every data set, with the intersection depending on which cutpoints classified the majority of controls. The rain threshold appears more important in situations where temperatures are just below or above minimum levels ( $\approx 270$  degree-hours) for an infection (25). The data suggested that other variables in the infection process were not accounted for when the 4-day degree-hour accumulation was high. *MARYBLTY* results showed a similar, but less drastic, effect. When comparing cutpoints 7 and 4 (temperature and EIP), with or without the rain threshold, the specificity of cutpoint 7 compared with 4 but sensitivity declined. Rain was important for identification of cases but further cases were missed, once again suggesting the need for additional variables (Figs. 2 and 5). Trauma events such as frost and high winds during bloom also have been implicated as causes of infection but were not considered in this study (29).

Nectar water potential and relative humidity >80% have been shown to play a significant role in the in vitro blossom infection process (21). High relative humidity, in particular, has been shown to be important to allow the total population of *E. amylovora* on pear stigmas to increase from  $10^6$  to  $10^7$  CFU per flower and also to allow multiplication in the hypanthium, where it generally was thought that little multiplication occurred (21,30). Relative humidity seems to be critical for multiplication of the bacteria to very high inoculum doses, as well as to supply enough moisture to stimulate substantial nectar production to allow for multiplication and possible ingress of *E. amylovora*.

The effect of the rain threshold also is variable among geographic regions. Although the rain threshold did not significantly improve *Cougarblight* in any geographic region, the shapes of the ROC curves changed (Fig. 5). With data from England, the shape of the two curves was more or less the same; however, for the east and west coasts of North America, cutpoint increases tended to increase only specificity, not sensitivity for *Cougarblight and rain*. This indicates that fewer noninfection events were misclassified but a similar number of infection events were identified. The most extreme example is found in the west coast data set, where the curve for *Cougarblight and rain* plateaus. *MARYBLTY* in the same data set also plateaus between cutpoints 3 and 7. It is unclear whether this was due to the rain threshold; however, both had very low AUCs, indicating that they were either only marginally better, or not better, than chance. The fact that both models with rain thresholds performed poorly suggests that the currently formulated rain thresholds were not adequately describ-

ing the blossom infection conditions in an arid climate. Factors such as whether or not irrigation is used in an orchard and the type of irrigation could change tree water potential drastically. The use of irrigation varies widely between geographic regions. Most apple orchards in British Columbia and Washington state use irrigation extensively, whereas many orchards in eastern North America and in England rely on natural rainfall. Where irrigation is used, the effect of tree water potential on infection would be reduced because the trees would rarely have nectar water potential too low for infection (21). It seems very unlikely that the irrigation system would be directly involved in wetting blossoms to initiate an infection event because most systems currently used are drip irrigators, keeping the water out of the canopy. The percent relative humidity in the arid interior valleys of the west coast states and British Columbia is substantially lower than in the more humid eastern climates. The rain threshold was important in the arid environment because it did identify noninfection events well. However, it may be that, in an arid climate, changes in the relative humidity that can occur due to irrigation or some other factor may play a larger role than first thought. It also is possible that some infection events were triggered by trauma such as frost or high winds but were indistinguishable from blossom blight in the data. These are areas that require further study, not only for improving blossom blight prediction systems but also for better general understanding of the infection process.

When Smith designed *Cougarblight* in the mid 1990s, one of his criticisms of available fire blight forecasting systems, including *MARYBLTY*, was the overprediction of infection events, contributing to the FPP (25). In an attempt to correct this problem, the 4-day degree-hour accumulation was used to approximate *E. amylovora* population growth, and an adjustment for inoculum dose, the potential for pathogen presence, was added (25). In the ROC curves (Fig. 5), the use of upper cutpoints in prediction of infection with both versions of *Cougarblight* gave high specificity (low FPP), much higher than *MARYBLTY* cutpoints in any data set. Whether this was due to the inoculum estimate or the degree-hour accumulation was unclear. However, the same *Cougarblight* cutpoints had relatively low sensitivity, mainly <0.5, showing that, although these cutpoints infrequently predicted a false infection event, actual infections were not predicted when needed. In contrast, the greatest specificity for any *MARYBLTY* cutpoint, because the forecaster is currently designed with 7 or fewer cutpoints, was 0.73. This is evidence of overprediction of infection events, substantiating the criticism of Smith (29). Although the specificity of *Cougarblight* was higher than *MARYBLTY*, the sensitivity of *MARYBLTY* cutpoints generally was higher, meaning that more infection events were detected. Overall, the predictions were about the same; however, *MARYBLTY* predictions were better at classifying cases and *Cougarblight* predictions at classifying controls.

Correlations between model predictions were calculated to see whether the models had similar predictions with the same data set. There was no consistent pattern among the various data sets where two of the models always had similar predictions (Table 4). Considering the overlapping 95% CIs of the AUCs for the models in each data set, it was expected that the correlation coefficient would be >0.80, as was reported in New York and Hungary (4,5). The highest correlation among model predictions occurred within the England data set. When looking at the cultivar data sets, it is interesting to note that although *MARYBLTY* and *Cougarblight and rain* had comparable forecast accuracy in the moderately susceptible cultivar data set but not with the very susceptible cultivars, the correlation coefficients were close in both data sets. *Cougarblight* predictions had low correlation coefficients with both *MARYBLTY* and *Cougarblight and rain*. With the cultivar data, *Cougarblight*, the only model without a rain threshold, identified separate infection events from the other two models.

A potential criticism of this study is that only 1 day per data set was used in the in ROC curve analysis. Because no quantitative symptom development information was available in the majority of the historical data sets, it was very difficult to determine which day with imminent infection risk was responsible for any observed infections. The actual conditions that cause infection could be important for refinement or development of thresholds to improve the accuracy of predictions. Even with symptom development data, it would be difficult to resolve whether more than one infection event was responsible, especially if there were several consecutive days that were highly conducive to infection. The potential problem of selecting a day as most likely for infection was that the rest of the predictions for that season were not used. No indication of how well the forecaster performed on a whole-season basis can be drawn from this ROC curve analysis. Many seasons had more than one predicted infection event; however, with disease incidence from historical data, there was no way to determine whether infection periods were under- or overpredicted within a season. A different method of assessment is needed to get the full picture of how the forecasters performed over the entire season. In other instances where ROC curve analysis has been used in assessing plant disease forecasters (11,32,33), the issue of multiple infection periods in a season has been either not a factor in the systems that were analyzed, or ignored.

An important question of the use of economic thresholds in plant disease management was raised in respect to ROC analysis by Hughes et al., Madden, Yuen, and Yuen and Hughes (11,14, 38,39). In many plant disease management systems, economic thresholds and an acceptable levels of crop loss are difficult to determine because of market variability. One manner of dealing with market variability is to develop ROC curves for several levels of damage. This leads to a different determination of cases and controls than is used in diagnostic medicine, and the implications are discussed in Hughes et al. (11). In this study, disease prevalence data were used rather than disease severity or an economic threshold, thereby defining cases and controls in the same manner as diagnostic medicine. This approach was taken for two reasons: (i) in many instances, incidence was the only disease measurement available in the historical data; and (ii) fire blight researchers, especially Steiner and Lightner (29), felt that any blossom infection was detrimental due to increased inoculum potential for later phases of the disease.

Economics also are an important consideration when determining the best cutpoint to use for a forecaster. Youden's index is commonly used for determining the overall non-error rate of a cutpoint. No single *Cougarblight* cutpoint could be identified as the overall most accurate with Youden's index (Table 5). The optimal cutpoint varied between the two forms of *Cougarblight* in the same data set; therefore, it is difficult to make a recommendation other than that the upper two cutpoints of each model tended to be the most accurate. The best cutpoint for *MARYBLYT* was determined to be number 7, imminent infection, in almost every situation. This showed that, if the TPP and FPP had equal importance in fire blight management, as was assumed with Youden's index, then the most accurate timing of preventative sprays would be at imminent infection (10,16,37). It should be noted that, although the fire blight forecasters in this study predicted blossom blight infection better than chance, the level of accuracy was relatively low, with the majority of Youden's index (Table 5) found to be <0.5 in this study. In the field, many consultants and growers also consider recommending a spray application at a *MARYBLYT* high risk level (H) in addition to imminent. This is because, in fire blight management, most people would prefer to have a false positive prediction, although not too frequently, thereby spraying when it is unnecessary. In addition, wetting events are difficult to predict and can be unevenly distributed, especially when showers are forecasted. A wetting event often is the factor missing from a high to an infection event. False

negative predictions are less desirable because of huge losses that could be incurred to replace an orchard due to a missed spray. The economic consequences of a missed spray are potentially much greater to an orchardist than applying an extra one (18). This ignores the long-term risk of the *E. amylovora* population developing antibiotic resistance in areas where it has not yet done so, such as populations have on the west coast of North America, and in Missouri and Michigan (13,15,23). To help balance all outcome consequences, cost function analysis of forecaster's cutpoints attempts to pick the best cutpoint when the cost of true positive predictions and false positive predictions are assumed unequal (10,16). A cost function analysis of the various decision outcomes currently is underway and will be used to better select the optimal *MARYBLYT* cutpoint for various disease prevalence levels.

It is possible that other fire blight forecasters, such as the Billing's revised system, are more accurate. However, it was decided that the two forecasters compared in this study were the most relevant in the context of fire blight forecasting in North America because they are the most commonly used (4,12,18, 25,34). It also was found that the Billing's revised system was difficult to interpret but, most importantly for this study, not easily computerized (2).

## ACKNOWLEDGMENTS

We thank the United States Department of Agriculture CREES North-east integrated pest management program for funding this project; all those who contributed historical data sets, making this work possible: E. Billing, L. Berkett, J. Charest, G. Jespersen, A. Jones, G. Lightner, V. Pillion, P. Sholberg, and T. Smith; T. Hall for his development of the Microsoft Excel-based version of *MARYBLYT*; and H. Aldwinckle and R. Seem for their helpful critiques and comments during the preparation of the manuscript.

## LITERATURE CITED

- Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* 12:387-415.
- Billing, E. 1992. Billing's revised system (BRS) for fireblight risk assessment. *EPPO Bull.* 22:1-102.
- Billing, E. 2000. Fire blight risk assessment systems and models. Pages 293-318 in: *Fire Blight: The Disease and Its Causative Agent, Erwinia amylovora*. J. L. Vanneste, ed. CAB International, Wallingford, UK.
- Breth, D. I., and Aldwinckle, H. S. 2002. Comparison of models for blossom blight prediction in New York. *Acta Hort.* 590:147-151.
- Bubán, T., Rutkai, E., Dorgai, L., and Thomson, S. V. 2004. Prediction infection risk on the basis of weather-related factors and *Erwinia amylovora* colonization in apple and pear flowers. *Int. J. Hortic. Sci.* 10:39-54.
- DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44:837-845.
- Dewdney, M. M., Biggs, A. R., Lightner, G. W., and Turechek, W. W. 2002. *MARYBLYT*: Can there be improvements? (Abstr.) *Phytopathology* 92(suppl.):S19.
- Dewdney, M. M., Biggs, A. R., and Turechek, W. W. 2003. A statistical comparison of *MARYBLYT* and *Cougarblight* using receiver operator characteristic (ROC) analysis. (Abstr.) *Phytopathology* 93(suppl.):S20.
- Hanley, J. A., and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29-36.
- Hughes, G., and Madden, L. V. 2003. Evaluating predictive models with application in regulatory policy for invasive weeds. *Agric. Syst.* 76:755-774.
- Hughes, G., McRoberts, N., and Burnett, F. J. 1999. Decision-making and diagnosis in disease management. *Plant Pathol.* 48:147-153.
- Jones, A. L. 1992. Evaluation of the computer model *MARYBLYT* for predicting fire blight blossom infection on apple in Michigan. *Plant Dis.* 76:344-347.
- Jones, A. L., and Schnabel, E. L. 2000. The development of streptomycin-resistant strains of *Erwinia amylovora*. Pages 235-251 in: *Fire Blight: The Disease and Its Causative Agent, Erwinia amylovora*. J. L. Vanneste, ed. CAB International, Wallingford, UK.
- Madden, L. V. 2006. Botanical epidemiology: Some key advances and its continuing role in disease management. *Eur. J. Plant Pathol.* 115:3-23.

15. McManus, P. S., and Stockwell, V. O. 2000. The use of antibiotics in agriculture: Silver bullet or rusty saber. APSnet Feature Story, June 2000. <http://www.apsnet.org/online/feature/Antibiotics>.
16. Metz, C. E. 1978. Basic principles of ROC analysis. *Semin. Nucl. Med.* VIII:283-298.
17. Murtaugh, P. A. 1996. The statistical evaluation of ecological indicators. *Ecol. Appl.* 6:132-139.
18. Norelli, J. L., Jones, A. L., and Aldwinckle, H. S. 2003. Fire blight management in the twenty-first century—using new technologies that enhance host resistance in apple. *Plant Dis.* 87:756-765.
19. Obuchowski, N. A., and Lieber, M. L. 1998. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Acad. Radiol.* 5:561-571.
20. Park, S. H., Goo, J. M., and Jo, C.-H. 2004. Receiver operating characteristic (ROC) curve: Practical review for radiologists. *Korean J Radiol.* 5:11-18.
21. Pusey, P. L. 2000. The role of water in epiphytic colonization and infection of pomaceous flowers by *Erwinia amylovora*. *Phytopathology* 90:1352-1357.
22. Schulzer, M. 1994. Diagnostic tests: A statistical review. *Muscle Nerve* 17:815-819.
23. Sholberg, P. L., Bedford, K. E., Haag, P., and Randall, P. 2001. Survey of *Erwinia amylovora* isolates from British Columbia for resistance to bactericides and virulence on apple. *Can. J. Plant Pathol.* 23:60-67.
24. Shtienberg, D., Shwartz, H., Oppenheim, D., Zilberstaine, M., Herzog, Z., Manulis, S., and Kritzman, G. 2003. Evaluation of local and imported fire blight warning systems in Israel. *Phytopathology* 93:356-363.
25. Smith, T. J. 1999. Report on the development and use of *Cougarblight* 98C—a situation-specific fire blight risk assessment model for apple and pear. *Acta Hort.* 489:429-436.
26. Smith, T. J. 2002. Fire blight daily risk assessment model. Version 2002C (Celsius). Published online by Washington State University. <http://www.ncw.wsu.edu/treefruit/fireblight/mdl98c.htm>.
27. Smith, T. J. 2002. Cougarblight 2002 Fire Blight Risk Assessment model. Published online by Washington State University. <http://www.ncw.wsu.edu/treefruit/fireblight/2000f.htm>.
28. Steiner, P. W. 1990. Predicting apple blossom infections by *Erwinia amylovora* using the *MARYBLYT* model. *Acta Hort.* 273:139-148.
29. Steiner, P. W., and Lightner, G. W. 1996. *MARYBLYT* 4.3: A Predictive Program for Forecasting Fire Blight Diseases in Apples and Pears. University of Maryland, College Park.
30. Thomson, S. V. 1986. The role of the stigma in fire blight infections. *Phytopathology* 76:476-482.
31. Thomson, S. V., and Gouk, S. C. 2003. Influence of age of apple flowers on growth of *Erwinia amylovora* and biological control agents. *Plant Dis.* 87:502-509.
32. Turechek, W. W., and Wilcox, W. F. 2005. Evaluating predictors of apple scab with receiver operating characteristic curve analysis. *Phytopathology* 95:679-691.
33. Twengström, E., Sigvald, R., Svensson, C., and Yuen, J. 1998. Forecasting *Sclerotinia* stem rot in spring sown oilseed rape. *Crop Prot.* 17:405-411.
34. van der Zwet, T., and Beer, S. V. 1999. Fire blight—its nature, prevention, and control. A practical guide to integrated disease management. *AIB* 631:1-91.
35. van der Zwet, T., Biggs, A. R., Heflebower, R., and Lightner, G. W. 1994. Evaluation of the *MARYBLYT* computer model for predicting blossom blight on apple in West Virginia and Maryland. *Plant Dis.* 78:225-230.
36. Vida, S. *AccuROC for Windows 95/98/NT Version 2.5*. 2001. McGill University Health Center, Montréal, PQ, Canada.
37. Youden, W. J. 1950. Index for rating diagnostic tests. *Cancer* 3:32-35.
38. Yuen, J. 2003. Bayesian approaches to plant disease forecasting. *Plant Health Progress Online*: doi:10.1094/PHP-2003-1113-06-RV.
39. Yuen, J., and Hughes, G. 2002. Bayesian analysis of plant disease prediction. *Plant Pathol.* 51:407-412.