Discovering Risk of Disease with a Learning Classifier System

John H. Holmes
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
Philadelphia, PA 19104
holmes@cceb.med.upenn.edu

Abstract

A learning classifier system, EpiCS, was used to derive a continuous measure of disease risk in a series of 250 individuals. Using the area under the receiver-operating characteristic curve, this measure was compared with the risk estimate derived for the same individuals by logistic regression. Over 20 training-testing trials, risk estimates derived by EpiCS were consistently more accurate (mean area=0.97, SD=0.01) than that derived by logistic regression (mean area=0.89, SD=0.02). The areas for the trials with minimum and maximum classification performance on testing were significantly greater ($p=0.019$ and $p<0.001$, respectively) than the area for the logistic regression curve. This investigation demonstrated the ability of a learning classifier system to produce output that is clinically meaningful in diagnostic classification.

1.0 INTRODUCTION

This work investigated the use of a learning classifier system to discover a type of knowledge particularly useful to epidemiologic researchers and clinicians: risk of disease. The focus of this investigation was twofold. First, it sought to refine and evaluate a learning classifier system based on NEWBOOLE (Bonelli et al 1991), re-engineered to function in clinical problem domains. Second, this work sought to use this system to identify factors in individuals with a given disease, such that one could use these factors in diagnosing new patients. Specifically, match set statistics were evaluated as a method for deriving disease risk. This introduction discusses several key concepts in clinical epidemiology, such as risk and clinical decision rules, and especially the area under the receiver-operating characteristic curve, a method for evaluating the accuracy of clinical tests and machine learning systems.

1.1 RISK

An essential goal of epidemiologic research is the elicitation of cause and effect relationships, or *causation*. Traditionally, epidemiologists (Rothman 1986) and classical philosophers (Hume 1739) have avoided the notion of absolute causation, that an event or other factor can be shown to cause another, because of the possibility of coincidence (Hume's problem), incorrect sampling, or some other flaw inherent in observation. Sample-based statistical methods of inference allow the epidemiologist to avoid gracefully the problem of defending causation. Rather than assert that a factor *X* (such as a toxin exposure) will *cause* disease *Y*, epidemiologists are more comfortable in saying that individuals with exposure to factor *X* are at increased *risk* of developing disease *Y*. Risk is the probability of occurrence of some event; thus, risk is often expressed in terms of chance: "Individual A has a 30% higher chance of contracting a pneumococcal infection than Individual B." A risk factor is some condition, event, disease, or other characteristic which is statistically associated with an increase in risk for a given disease. Frequently, the outcome of interest in an epidemiologic investigation is binary in value, such as diseased/non-diseased, or dead/alive. In addition, it is important to adjust for a number of factors that may affect the risk associated with a given factor; this adjustment is performed by multivariate modeling.

When performing multivariate statistical analysis where a binary outcome variable is used, the preferred method is logistic regression, in which the risk of developing an outcome is expressed as a function of a set of predictor (or independent) variables. The dependent variable in the logistic model is the natural logarithm of the odds of disease:

$$\ln\left[\frac{P_x}{1-P_x}\right] = \alpha + b_1 X_1 + \ldots + b_n X_n \qquad (1)$$

where P_x is the probability of disease for a specified covariate x, $\ln\left[\frac{Px}{1-Px}\right]$ is the log odds of developing the outcome (or logit), accounting for all exposure variables in the model (rather than a single exposure variable), a is the intercept, and b_i is the coefficient for the independent variable X_i.. Rewriting this equation produces a method for representing the estimated probability of developing the outcome of interest:

$$P\hat{y} = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \ldots + \beta_n x_n)}} \qquad (2)$$

where $\hat{y}$ is the estimated probability of developing an outcome, given the presence of risk factors $x_1 \ldots x_n$. The formula in Equation 2 is used to derive clinical decision rules.

1.2 CLINICAL DECISION RULES

Another important goal of epidemiology is to develop rules (actually complexes of rules and concepts) that can be used to classify individuals as being at risk for having a disease or other clinical outcome. *Clinical decision rules* are used to classify patients in terms of their risk of developing a specific outcome. While the ultimate goal of applying a decision rule is to determine a categorical diagnostic classification into which a patient will fall (such as "Disease" or "No Disease"), the categories are actually created from a continuous measure of risk derived from Equation 2.

Once decision rules are created, they must be validated, usually through testing in other samples of the population from which they were derived. A number of measures exist for validating decision rules as well as other diagnostic tools.

1.3 DIAGNOSTIC TEST EVALUATION

In order to determine the validity of a test or procedure which classifies a dichotomous outcome, clinical epidemiologists employ the 2x2 contingency table. An example is shown in Figure 1, which is a 2x2 table of white blood cell count (WBC), determined by microscopic examination, and infection, determined by blood culture as the "gold standard", or the true indicator of the presence of disease. In this example, the WBC was considered to be positive (diagnostic of infection) if it exceeded 5,000.

WBC	Infection	
	Positive	Negative
Positive	95	72
Negative	5	28

Figure 1. A 2x2 contingency table for a hypothetical study of white blood cell count and diagnosis of infection in 200 individuals

From the four cells, several measures can be calculated, each describing some aspect of the classification or predictive validity of the test. All of these measures are proportions, ranging in value from 0.0 to 1.0. Two such measures, *sensitivity* and *specificity*, are of interest here.

1.3.1 Sensitivity and specificity

Sensitivity indicates a test's ability to classify correctly individuals actually having the disease. Sensitivity is analogous to the *true positive rate*. Thus, from Figure 1,

$$\text{Sensitivity} = \frac{95}{100} = 0.95$$

indicating that a WBC of 5,000 will correctly classify 95% of those patients who are actually infected.

Specificity measures the ability of a test to classify correctly those without disease; specificity is also called the *true negative rate*. The specificity of the WBC (from Figure 1) is:

$$\text{Specificity} = \frac{28}{100} = 0.28$$

Thus, the WBC will correctly classify 28% of those truly non-infected patients. Another way of looking at sensitivity is the *false positive rate*, or the quantity (1-specificity); this value (0.72 in the example) shows that 72% of patients with a WBC of 5,000 will be classified incorrectly as positive for infection.

The relationship between the true positive and false positive rates demonstrates the amount of information in a test, and can be demonstrated graphically on a *receiver-operating characteristic curve.*

1.3.2 The Receiver-Operating Characteristic (ROC) Curve

1.3.2.1 Introduction

The ROC curve is created by plotting the true positive rate (sensitivity) on the vertical axis against the false positive rate (1-specificity) on the horizontal axis. An ROC curve is ordinarily plotted over a range of values, or *cutoffs*, for a given diagnostic test. For diagnostic tests that are expressed as continuous values these cutoffs represent threshold values at which a patient would be classified as positive; consequently, the correct determination of these cutoffs represents an important problem in epidemiologic research. The table

in Figure 2 shows the sensitivities and specificities associated with a range of cutoffs for WBC.

WBC	Sensitivity	Specificity
5,000	0.95	0.28
10,000	0.88	0.52
15,000	0.60	0.88
20,000	0.40	0.93

Figure 2. Cutoff values of white blood cell count in diagnosing infection.

Using separate 2x2 contingency tables created at each cutoff value, one can evaluate the effect of broadening or narrowing the diagnostic criterion (the cutoff) on the sensitivity and specificity of the test. Figure 2 demonstrates that requiring a cutoff WBC of 20,000 to diagnose infection would result in very few false positives (specificity=0.93, or a false positive rate of 0.07). However, the sensitivity (0.40) at this cutoff indicates that many patients with infection would be missed, if a WBC of 20,000 or greater were required to make a positive diagnosis. Using a cutoff of 5,000 would result in fewer patients with infection being missed (sensitivity=0.95), but also in the misdiagnosis of many patients truly without infection (specificity=0.28). Using the values from Figure 2, the ROC curve shown below in Figure 3 can be drawn.

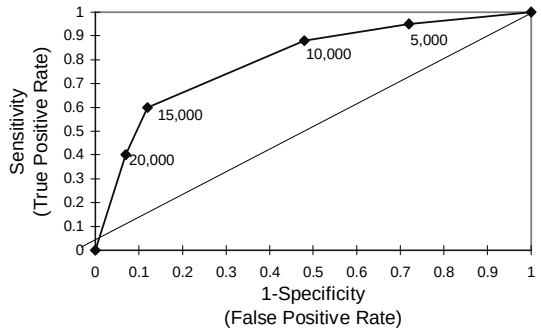

Figure 3. Receiver-operating characteristic curve for determining presence of infection based on cutoffs of white blood cell count

The optimal cutoff value could be selected based on a visual inspection of the curve. For example, one could choose a cutoff of 20,000 which would result in a very low false positive rate, but at the expense of sensitivity. A better cutoff would be 15,000 since the false positive rate remains low, and the sensitivity increases to 0.60.

The ROC curve is also employed to determine the overall usefulness of a diagnostic test. In order to determine whether a test is useful , it must be evaluated for its discrimination accuracy, or its ability to classify normal and abnormal patients. The measure used for this purpose is the area under the ROC curve.

1.3.2.2 The area under the ROC curve (θ)

While the graphical representation of the ROC curve is of interest in determining appropriate diagnostic cutoffs for a given test, the area under the ROC curve is important for demonstrating the ability of the test to classify both true positives and true negatives, simultaneously, as a single measure. The area under the ROC curve has been used extensively in medical decision making as a standard method for evaluating diagnostic test performance (Coultrip and Grossman 1993; Almog et al 1992; Cowan et al 1992; Pahlm et al 1990; Good et al 1990; Van Lente et al 1990; Somoza,et al 1990).

The area under the ROC curve is expressed as a proportion of the total possible area defined by the x- and y-axis. The area represents the probability of a true response in a two-alternative forced-choice (Yes-No) task; thus, the quantity (1-area) is the false alarm rate (Green and Swets 1966). The ROC curve for a test which classifies well will have a "shoulder" closer to the upper left-hand corner of the plot, and farther from the 45-degree diagonal; this test would have a high sensitivity and a high specificity and, as a result, a higher area. A test which contains no information would plot on the 45-degree diagonal through the origin, because the true positive rate (sensitivity) and false positive rate (1-specificity) would be equal across all cutoffs.. The area under such a “curve” would be 0.50, indicating that the test would discriminate only as well as a coin-flip; that is, there would be a 50% chance of error (or “false alarm”) when such a test is used. A nonparametric method based on the Wilcoxson statistic (W) and its standard error (SE_W) is regularly used in approximating the area under the ROC curve, θ, and its standard error, SE_θ (Hanley and McNeil 1982).

The ROC curves of several tests can be compared and their areas evaluated for statistically significant differences to determine which test is best in overall discrimination ability. This is done by performing a Z test on the nonparametric areas (Hanley and McNeil 1982; McNeil and Hanley 1984). If the area of one test is not significantly different than another, then both tests are considered to be equally “good” in terms of predictive ability. However, if there is a significant difference between their areas, the test with the higher area is considered to be the better predictor of the two.

This investigation employed the area under the ROC curve in evaluating the classification ability of a learning classifier system, and the ability of such a system to derive estimates of disease risk.

1.4 FOCUS OF THIS STUDY

This investigation evaluated the ability of a learning classifier system to assign an estimate of risk of outcome to a novel, previously unseen observation, given a specific combination of features in that observation's taxon. This differs from the usual approach of assigning a single nominal-level classification to a novel case. Thus, a different paradigm was used to determine the type of decision made by the system; this was based upon the proportions of positives and negatives in the *match set* (the set of classifiers in a population matching the taxon of an input case).

2.0 METHODS

2.1 TESTBED DATA

Sham datasets consisting of 15 demographic, medical history, and exposure variables, one outcome variable, and 500 observations were created using the random data generator routines supplied with the EpiInfo (Dean et al 1990) epidemiologic analysis software package. The datasets represented epidemiologic surveillance for hepatocellular (liver) carcinoma in a group of individuals living in within the same neighborhood who may have been exposed to one or more suspected hepatocarcinogens. All variables were coded dichotomously, with 0s or 1s used to indicate the absence or presence, respectively, or value categories, of a variable.

The dataset was created in such a way as to bias one variable (exposure to Vinyl Chloride (ViCl), a known hepatocarcinogen) toward association with the outcome. All other variables were randomly assigned a value of 0 or 1 using a random normal deviate procedure. Training and testing sets were created by randomly selecting records the dataset at a sampling fraction of 0.50 without replacement; thus, training and testing sets were equal in size, equal in number of positive and negative examples, and mutually exclusive.

2.2 TRAINING-TESTING SEQUENCE

A total of 20 trials, each consisting of a training period and a testing epoch, were performed. During the training epoch, cases selected in random order from the training set were presented to the system over a total of 30,000 iterations; a single case presentation comprised one iteration. The system was evaluated by calculating the area under the ROC curve during training ($\theta_{training}$) at every 100th iteration to monitor learning performance. This was achieved by testing the system with every case in the training set. At the conclusion of training, the system was tested with each case in the testing set, and its classification performance monitored by calculating the area under the ROC curve ($\theta_{testing}$). The results reported here focus on the testing epochs of the trials with the minimum and maximum $\theta_{testing}$ as well as on an average across all 20 trials.

2.3 EpiCS: TESTBED LEARNING CLASSIFIER SYSTEM

An object-oriented version of NEWBOOLE (Bonelli et al (1990), called EpiCS, was created and used as the classifier system in this investigation. EpiCS departed from NEWBOOLE and its predecessor, BOOLE (Wilson 1987) on several features: population size, algorithms for controlling under- and over-generalization, and a methodology for determining risk as a measure of classification.

Population size

The testbed data required 15 positions on a taxon (rather than the six needed for the 6-multiplexer). It is generally accepted that longer strings will require larger populations to prevent population overcrowding with overly specific strings and to improve system performance. Although a method has been proposed for calculating the optimal population size for genetic algorithms (Goldberg 1989), the best heuristic for determining population size for classifier systems is that "more is better" (Robertson and Riolo 1988). Given this heuristic, and given that NEWBOOLE was parameterized at a population size of 1,000 for the 11-multiplexer problem, the performance of EpiCS using several population sizes between 400 and 2,000 was investigated. It was found that population sizes over 1,000 caused serious degradation in system performance, proportional to increasing population sizes. In addition, little improvement in learning rate was noted with increasing population size. Population sizes less than 800 resulted in overly general populations of classifiers, likely owing to the generalization pressure described by Robertson and Riolo (1988). Increasing the population size from 800 to 1,000 improved the learning rate and significantly decreased overgeneralization. As a result, the population size for the investigation of epidemiologic surveillance data was fixed at 1,000.

Controlling over-generalization

Epidemiologic data are generally noisy, or full of contradictions between exposure factors and outcomes. For example, not everyone exposed to a

particular factor will develop a given disease. Early investigations with EpiCS showed that the system tended to overgeneralize the population when trained with noisy data, such that the classifications made by the system were useless. In these situations, the classifiers would be sufficiently strong to survive training, but would still contain a surfeit of non-specific bits (*s). Robertson and Riolo (1988) suggested that overgeneralization can be controlled in classifier systems, via taxation or classifier deletion based on time since last use. In this investigation, a third method was employed to control overgeneralization in the population. A *governor* was implemented in EpiCS, which evaluated each classifier in the population at each iteration, and re-initialized overly general classifiers with randomly-assigned bits. The effect of this procedure was to delete overly general classifiers and replace them with less general classifiers. After experimentation with a range of values from 0.50 to 1.00, the governor was ultimately parameterized at 0.85; thus, a classifier with *s comprising more than 85% of its length would be re-initialized.

Controlling under-generalization

In using EpiCS on noisy data, it was found that a significant proportion (10%) of testing cases could not be classified, correctly or incorrectly. Examination of the classifier population at the end of training revealed that it had a preponderance of highly specific classifiers. As a result, the classifier population at the end of training was ill-equipped to classify testing cases, and this resulted in poor classification performance. As does its predecessor NEWBOOLE, EpiCS employs a penalty factor, *p*, which penalizes classifiers advocating incorrect decisions. Experimenting with the penalty factor *p* over a range of values (0.25, 0.50, and 0.75) led to the observation that the value of *p* originally used in NEWBOOLE (0.95) was apparently too high, resulting in the premature demise of general (but useful) classifiers. It was found that decreasing *p* to 0.50 resulted in much-improved performance and decreased numbers of unclassifiable cases on testing.

Determination of risk

While match sets are used during the training and testing epochs, this investigation focused on those created during the evaluation of testing cases.

For example, a case in the training set would be labeled as a positive if, in the match set, there existed a preponderance of classifiers predicting positive for disease. This would be the situation even if the strongest classifier in the match set predicted a negative outcome. Thus, rather than produce an output decision based upon strength, the system would predict an outcome based upon the *prevalent* outcome present in the match set. The output of the system would now be *risk of disease*, rather than a dichotomous decision (that being the presence or absence of disease). The advantage of this approach is that it provides a decision which is continuous in nature and which can be cut at various points along its range.

To implement this risk-based classification paradigm, each case in the testing set was presented to the trained classifier system and evaluated for the probabilities of presence and absence of disease; these were determined from the proportion of classifiers matching a given input case taxon. For example, suppose an input case (reduced for purposes of illustration to a five-bit taxon, 01100) were presented to the system. Further suppose that four candidate classifiers (those with taxa matching that of the input taxon) were extant in the match set, each with a probability (P) based on proportionate representation in the match set:

Classifiers	Proportion in match set
*10*0:1	0.60
0*100:0	0.05
01*00:1	0.22
0***0:1	0.13

Assuming that the action bit represented the presence of disease with a 1 and the absence with a 0, the classifier system-derived probability of disease (CSPD) for the given input case would be 0.95:

$$\text{CSPD} = \frac{\sum(\text{P(disease - positive classifiers)})}{\sum(\text{P(matching classifiers)})} = 0.95 \quad (3)$$

This procedure was repeated for each case in the testing set at each prevalence.

In order to compare the CSPD with an accepted method of determining the probability of disease, the logistic regression-derived probability of disease (LRPD), or risk estimate, was calculated for each of the testing cases for using a clinical decision rule derived from the training cases in the appropriate dataset. It was shown above (Equation 2) that probability of disease can be estimated from a decision rule derived from a logistic model. To derive the decision rule, non-stepwise logistic regression was run on the training set, using diagnosis of hepatocellular carcinoma as the dependent variable and the 15 history and exposure variables as independent terms.

The decision rule was then applied to each case in the testing set in order to obtain the LRPD for each testing case. The application of the decision rule to obtain risk estimates on a sample individual is

demonstrated in Figure 4. Applying these values to the formula in Equation 4 yields a very high *Py*; with the LRPD, the individual in this example has a 99.4% chance of developing hepatocellular carcinoma, given her pattern of history and exposures.

The CSPD and LRPD were categorized by deciles to reflect specific disease probability cutoffs from 0 through 1, and true and false positive rates were calculated at each cutoff. Categorizing the probabilities of disease obtained from EpiCS and from logistic regression provided the means for comparing them using ROC curve analysis. Three ROC curves were constructed: one for the CSPD obtained from the trial with the minimum $\theta_{testing}$ ($CSPD_{min}$), one for CSPD from the trial with the maximum $\theta_{testing}$, ($CSPD_{max}$,) and one for LRPD. The areas under the two CSPD curves were compared with that under the LRPD curve for significant difference using the nonparametric method (Hanley and McNeil 1982; McNeil and Hanley 1984). It was necessary for the θ of both $CSPD_{min}$ and $CSPD_{max}$ to be significantly greater than the θ of the LRPD in order to conclude that the classifier system method of deriving risk was superior to the logistic regression method. If there were no significant difference between the two curves, it could be assumed that the two methods were equally good (or bad) at patient classification.

Variable	Coefficient	Value of variable	Coefficient*Value
Intercept	-9.8303	1	-9.8303
Age	0.7744	1 (>=50 years old)	0.7744
Sex	-0.0664	1 (Female)	-0.0664
History of cancer	0.7053	0 (no)	0
History of cirrhosis	0.7111	0 (no)	0
Exposure to vinyl chloride	8.9072	1 (yes)	8.9072
History of hepatitis C	4.1089	1 (yes)	4.1089
History of heavy alcohol use	1.2347	0 (no)	0
History of drug abuse	1.2563	0 (no)	0
History of AIDS	1.5065	0 (no)	0
History of HIV seropositivity	-0.6251	0 (no)	0
History of cigarette smoking	1.3035	1 (yes)	0.013
Family history of cancer	-0.8412	0 (no)	0
Exposure to PCBs	0.1945	0 (no)	0
Exposure to methylmercury	0.8624	0 (no)	0
Total time on payroll	1.1775	1 (yes)	1.1775
		Total	5.0843

Figure 4. Application of a decision rule derived from a logistic regression model to a sample individual.

$$LRPD = Py = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\ldots+\beta_n x_n)}} = \frac{1}{1+e^{-(5.0843)}} = 0.994 \qquad (4)$$

3.0 RESULTS

Figure 5 shows the ROC curves obtained from the EpiCS trials associated with the minimum and maximum $\theta_{testing}$, compared with that obtained from logistic regression. The individual data points represent cutoff probabilities in 0.10 increments. Visual inspection of these curves reveals that EpiCS classified novel cases consistently better than did a decision rule derived from logistic regression. This is especially evident in comparing the cutoffs. The cutoffs for the EpiCS curves are optimal at 0.10; that is, a very high sensitivity (>=0.90) is obtained at a very low false positive rate (<0.10). This indicates that the classification accuracy of EpiCS is both highly sensitive and highly specific, even at very low probability of disease (0.10). The curve for logistic regression shows that this method does not approximate the classification accuracy of EpiCS until a cutoff of disease probability at 0.60 is used.

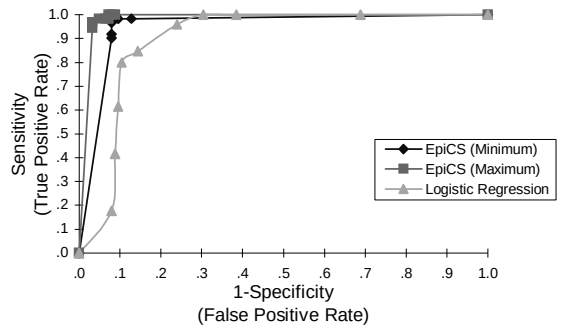

Figure 5. ROC curves obtained for the disease-positive risk estimates derived from EpiCS at minimum and maximum classification accuracy on testing, compared with the ROC curve obtained from logistic regression.

Figure 6 quantifies the degree of difference between the performance of EpiCS and logistic regression. The θs of both EpiCS curves were significantly greater (p=0.019 and p<0.001, respectively, for the minimum and maximum trials) than the θ of the logistic regression curve.

	θ	SE θ	p-value
Logistic Regression	0.90	0.02	---
EpiCS Minimum	0.95	0.02	0.019
EpiCS Maximum	0.98	0.01	<0.001

Figure 6. Areas under the ROC curves obtained from the EpiCS trials having the minimum and maximum θ on testing, compared with area under the ROC curve derived from logistic regression.

Over the 20 training-testing trials, the mean θ for EpiCS was 0.97 (SD=0.01), compared with the θ derived from logistic regression, 0.89 (SE=0.02).

4.0 DISCUSSION AND CONCLUSION

This investigation compared the classification performance of EpiCS with that of a traditional statistical method for classification, logistic regression. First, EpiCS was found to perform better than logistic regression in deriving an estimate of risk of disease for noisy, epidemiologic data. Excellent classification ability was noted for EpiCS over all training-testing trials.

Second, this investigation demonstrated the successful implementation of risk-based output from a learning classifier system, rather than a single, categorical classification. This investigation further demonstrated that the match set could be exploited during the testing epoch to assign a continuous estimate of risk that could, in turn, be employed for constructing ROC curves over a range of risk cutoffs. The primary advantage of this approach was shown here: the comparison of the areas under two ROC curves. In addition, each curve could be examined to determine the optimal risk cutoff for each method. While not a focus of this investigation, the ability to optimize human decisions made with the assistance of EpiCS would be a fertile area for additional investigation.

Third, logistic regression, like all inferential statistical methods, relies on sufficient numbers of cases to be valid. In addition, these statistical methods rely on numerous assumptions as to the characteristics of the data. In the early phases of epidemics, or in the case of rare diseases, where small numbers of disease-positive cases are available, it is often difficult to meet these assumptions. However, researchers still need a method for characterizing diseased and non-diseased individuals so that those at risk may be identified. A learning classifier system such as EpiCS fills this need because it is not constrained by the assumptions that may hamper traditional statistical analysis.

Finally, epidemiologists and practicing clinicians are seldom content to accept a single categorical classification for a given patient. They are much more comfortable with probabilities of outcomes, given specific patterns of features, so they can frame a patient's chances of having a disease along a continuous scale, rather than a simple yes or no classification. The ability of a classifier system to provide a continuous measure of risk is paramount to the goal of implementing genetics-based machine learning systems in clinical domains.

Acknowledgment

The author gratefully acknowledges Stewart W. Wilson, PhD, of the Rowland Institute for Science for his insightful comments on this work.

References

Almog, R; Goldkrand, JW; Saulsbery RA; Samsonoff C. Prediction of respiratory distress syndrome by a new colorimetric assay. Am J Obstet Gynecol. 1992 Jun; 166(6 Pt 1): 1827-32.

Bonelli, P.; Parodi, A.; Sen, S.; Wilson, S. NEWBOOLE: A fast GBML system. Porter, B.; Mooney, R. Machine Learning: Proceedings of the Seventh International Conference; 1990 Jun 21; Austin, Texas. San Mateo, CA: Morgan Kaufmann Publishers, Inc.; 1990: 153-159.

Coultrip, LL.; Grossman, JH. Evaluation of rapid diagnostic tests in the detection of microbial invasion of the amniotic cavity. Am J Obstet Gynecol. 1992 Nov; 167(5): 1231-42.

Cowan, BD.; Vandermolen, DT.; Long, CA.; Whitworth N S. Receiver-operating characteristic, efficiency analysis, and predictive value of serum progesterone concentration as a test for abnormal gestations. Am J Obstet Gynecol. 1992 Jun; 166(6 Pt 1): 1729-34.

Dean, AD; Dean, JA; Burton, JH; Dicker, RC. Epi Info, Version 5: a word processing, database, and statistics program for epidemiology on microcomputers. Centers for Disease Control, Atlanta, Georgia, 1990.

Goldberg, DE. Genetic Algorithms in Search, Optimization, and Machine Learning. New York: Addison-Wesley; 1989.

Good, WF; Gur, D; Straub, WH.; Feist, JH. Comparing imaging systems by ROC studies. Detection versus interpretation. Invest Radiol. 1989 Nov; 24(11): 932-3.

Green, DM; Swets, JA. Signal Detection Theory and Psychophysics. New York: John Wiley Sons; 1966.

Hanley, JA; McNeil, BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 143:29-36.

Hume, DA. A Treatise of Human Nature (1739). Second edition. Oxford: Clarendon Press; 1978.

McNeil, BJ; Hanley, JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making. 1984; 4:137-150.

Pahlm, O; Case, D; Howard, G; Pope, J; Haisty, WK. Decision rules for the ECG diagnosis of inferior myocardial infarction. Comput Biomed Res. 1990 Aug; 23(4): 332-45.

Robertson, GG; Riolo, RL. A tale of two classifier systems. Machine Learning. 1988; 3:139-159.

Rothman, KJ. Modern Epidemiology. Boston: Little, Brown and Company, 1986.

Somoza, E.; Soutullo-Esperon, L.; Mossman, D. Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory. Int J Biomed Comput. 1989 Sep; 24(3): 153-89.

Van Lente, F; Martin, A; Ratliff, NB; Kazmierczak, SC.; Loop, FD. The predictive value of serum enzymes for perioperative myocardial infarction after cardiac operations. An autopsy study. J Thorac Cardiovasc Surg. 1990 Nov; 98(5 Pt 1): 704-10.

Wilson, SW. Classifier systems and the animat problem. Machine Learning. 1987; 2: 199-228.