Reprinted from *Computers in Biomedical Research*,**26:** 74-97, 1993

A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain*

William J. Long,
MIT Laboratory for Computer Science, Cambridge, MA, USA
John L. Griffith and Harry P. Selker,
Center for Cardiovascular Health Services Research,
New England Medical Center, Boston, MA, USA
and Ralph B. D'Agostino
Mathematics Department, Boston University, Boston, MA, USA

Abstract

This paper compares the performance of logistic regression to decision-tree induction in classifying patients as having acute cardiac ischemia. This comparison was performed using the database of 5,773 patients originally used to develop the logistic-regression tool and test it prospectively. Both the ability to classify cases and ability to estimate the probability of ischemia were compared on the default tree generated by the C4 version of ID3. They were also compared on a tree optimized on the learning set by increased pruning of overspecified branches, and on a tree incorporating clinical considerations. Both the LR tool and the improved trees performed at a level fairly close to that of the physicians, although the LR tool definitely performed better than the decision tree. There were a number of differences in the performance of the two methods, shedding light on their strengths and weaknesses.

1 Introduction

The use of data to develop decision procedures has a long history with different approaches developed in different research communities. In the statistics community, logistic regression (LR) has assumed a major position as a method for predicting outcomes based on the specific features of an individual case. In the machine-learning community, techniques for decision-tree induction have been developed by a number of researchers. Interestingly, decision-tree induction techniques have also been developed in the statistics community, but have been called "regression trees" there.

These two techniques, logistic regression and decision-tree induction have often been used for very similar tasks. Pozen et al[1, 2], and Selker et al[3], use LR to develop predictive instruments for determining the probability that an emergency-room patient with chest pain or other related symptoms actually has acute cardiac ischemia. Goldman et al[4, 5], used the CART (Classification and Regression Trees) methodology[6] to develop decision trees for deciding whether patients

*This research was supported by National Institutes of Health Grant No. R01 HL33041 from the National Heart, Lung, and Blood Institute, No. R01 LM04493 from the National Library of Medicine, and R01 HS02068, R01 HS05549 and R01 HS06208 from the Agency for Health Care Policy and Research.

entering the emergency room with acute chest pain should be admitted to rule out myocardial infarction.

Given that both approaches are being used for similar purposes, it is important to gain an understanding of the relationship between statistical regression techniques and decision-tree techniques, and their relative strengths and weaknesses. A few papers have started to look at this issue. Mingers[7] compared the ID3 rule induction algorithm (using the G-statistic rather than Quinlan's information measure) to multiple regression on a data base of football results using 164 games in the learning set and 182 games in the test set. The results of this comparison favor ID3, but it is hard to draw any general conclusions because of the rather artificial nature of the five variables derived from past scores, only two of which were used by the multiple regression. Segal and Bloch[8] compared proportional-hazard models to regression trees in two studies of 604 patients and 435 patients respectively. The variables selected by each method were similar but beyond that, they were difficult to compare. The conclusion was that they are complementary methodologies, since each can provide insights not available with the other. They suggest using the selection of nodes in the tree to suggest variables and interaction terms for the equation and also using the regression statistical test (in this case Cox partial likelihood) as the splitting criterion for generating trees. Harrell et al[9] compared CART trees to several other strategies. Stepwise regression did not perform as well as CART on a training sample of 110, but better than CART on a training sample of 224. Kinney and Murphy[10] compared ID3 to discriminant analysis in a learning set with 107 items and a test set of 67 items in a medical domain (detecting aortic regurgitation by auscultation). Their conclusion was that both methods performed equally poorly. A more recent study, Gilpin, et al[11], compared regression trees, stepwise linear discriminant analysis, logistic regression, and three cardiologists predicting the probability of one-year survival of patients who had myocardial infarctions. This comparison used 781 patients in the learning set and 400 in the test set. The various methods were compared as to their sensitivity, specificity, accuracy, and area under the ROC curve. No statistically significant differences were noted between any of the methods. The area under the ROC curve for the regression tree was less than for the two multivariate equations, but the generated tree only had five leaves and therefore there were only five points on the ROC curve. At the selected probability breakpoint, the accuracy of the tree was between that of the two multivariate equations.

Since all of these tests were done on datasets of size in the hundreds, a possible reason for the general lack of significant differences between the methodologies is the small size. In the following study we will use the data set collected by Pozen and colleagues[2] containing 5,773 patients. This offers an opportunity to compare logistic regression and decision-tree induction on a large dataset where the existing logistic regression equation was carefully prepared and thoroughly tested.

2 Methodology

2.1 Decision-Tree Generation

Techniques for generating decision trees, also called classification or regression trees[6], have been developed over the past twenty years. In the machine-learning community, a number of researchers have been developing methods for inducing decision trees automatically from data sets. The best known of these methods are AQ11[12] and ID3[13], each of which has spawned a family of programs to fit the demands of real world domains. In the statistics community the CART program is the best known of the programs. Current versions of these programs are able to handle noisy data, continuous variables, variables with multiple (more than two) values, missing data, and classes with multiple values.

The nature of the task for a decision-tree program is as follows: The data set consists of a set of *objects*, each of which belongs to a *class*. There is a set of attributes (also called variables) with each object having *values* for the attributes. The task is to use the attributes to find a *decision tree* that classifies the data appropriately. Since there is a large number of such trees, the task is refined to find a small tree that classifies the data appropriately. If the set includes "noisy data" (variables whose values may not always be correct) or if the attributes are not always sufficient to classify the data correctly, the tree should only include branches that are justified adequately.

All of these programs work by recursively picking the attribute that provides the best classification of the remaining subset. That is, the program uses the whole data set to find the attribute that best classifies the data. Then for each subset defined by the values of the selected attribute, the process is repeated with the remaining attributes. Then the tree is pruned to eliminate branches that are not justified adequately. Thus, the two basic functions of a decision-tree program are 1) selecting the best attribute to divide a set at each branch, and 2) deciding whether each branch is justified adequately. The different decision-tree programs differ in how these are accomplished.

In the ID3 programs the best attribute is determined by computing the information gain ratio, derived from information theory. The information gain for an attribute A is (assuming p and n are positive and negative instances of a dichotomous classification variable):

$$G(A) = I(p,n) - \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

where

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

To handle continuous variables the program treats the variable as dichotomous by finding the cut point that maximizes the information gain. The definition is easily extended to variables with more than two values, but the information gain of a variable with several values is always greater than or equal to the information gain of the same variable with fewer values, even when the additional values do not contribute anything. To overcome this statistical bias, a normalization factor is introduced based on the information content of the value of the variable.

$$V(A) = -\sum_{i=1}^{v} \frac{p_i + n_i}{p+n}\log_2\frac{p_i + n_i}{p+n}$$

An alternative approach to multiple valued variables would be to look for a division of the values into two sets that maximizes the variable choice statistic. This is the approach taken by CART and is available in current versions of ID3, but is not used here since there is no need to lump the values together.

The information-gain-ratio statistic is not the only statistic that has been used to select attributes. Mingers[14] compares the information-gain statistic with the chi-square contingency table, the G statistic, probability calculations, the GINI index of diversity (the statistic used in the CART program[6]), the information-gain ratio, and the information gain with the Marshall correction. Since that paper was written, Quinlan and Rivest[15] have added the minimum description length principle to the armamentarium. Mingers' conclusion is that the predictive accuracy of the induced decision trees is not sensitive to the choice of statistic. This paper will use the C4 version of ID3[1] using the information-gain-ratio statistic, although this comparison could certainly be extended to include other statistics.

There are also multiple strategies for pruning the tree once it is generated. Quinlan[16] discusses five such strategies and finds the differences over a range of different kinds of data to be insignificant. The strategy used in C4 is called by Quinlan, pessimistic pruning. Normally a branch is pruned when the error introduced is within one standard error of the existing errors adjusted for the continuity correction. However, this paper will consider the number of standard errors to be a variable.

2.2 Logistic Regression

Logistic Regression is a non-linear regression technique that assumes that the expected probability of a dichotomous outcome is:

$$P = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots)}}$$

where the X_i are variables with numeric values (if dichotomous, they are, for example, zero for false and one for true) and the βs are the regression coefficients which quantify their contribution to the probability. Intuitively, the justification for this formula is that the log of the odds, a number that goes from $-\infty$ to $+\infty$, is a linear function. Given this model, stepwise selections of the variables can be made and the corresponding coefficients computed. In producing the LR equation, the maximum-likelihood ratio is used to determine the statistical significance of the variables. Logistic Regression has proven to be very robust in a number of domains and proves an effective way of estimating probabilities from dichotomous variables.

A particularly attractive attribute of the original LR tool developed by Pozen, et al[1, 2], and the slightly modified version generated later by the same group[3] is that they have been shown to be useful clinically and are now being used in clinical settings to assist physician decision making.

3 Data for the Comparison

The data used for this comparison were collected for the multicenter development and clinical trial of the predictive instrument for coronary care unit admission[2]. The data were collected at six New England hospitals, ranging from urban teaching hospitals to rural non-teaching hospitals. The patients included all consenting adults (over age 40 for females or over age 30 for males) presenting in the emergency room with chief symptom of chest pain, jaw or left arm pain, shortness of breath, or changed pattern of angina pectoris. Data collection was conducted in two consecutive one year phases. The 2,801 patient descriptions available at the end of the first year were used to develop the LR equation. (An additional 652 patient descriptions collected while the LR equation was developed are included in the learning set for this paper, for a total of 3,453 patients.)

The LR equation was developed from 59 clinical features available in the emergency room, including clinical presentation, history, physical findings, electrocardiogram, sociodemographic characteristics, and coronary disease risk-factors. The final equation only uses seven of the variables. The remaining clinical features were rejected for modeling reasons, and the interest of parsimony,

[1] Since the generic name ID3 is more widely known than the specific C4 version, the paper use the name ID3.

since a model using less than ten variables was desired. The final LR equation provides an estimate of the probability that the patient has acute cardiac ischemia. The test set of 2,320 was collected in the same way in a second year at the same hospitals, as part of a controlled trial conducted to assess the impact of providing physicians with the probability of acute ischemia determined by the LR equation. The assignment of the final diagnosis was done by blinded expert-physician reviewers and therefore was not affected by the intervention. All patients had the same follow-up as obtained in the first year.

To generate the decision trees all of the variables were used except three (related to the means of transportation used to the emergency room) that were not generalizable. Some of the variables used for the LR analysis (eg, *chief complaint chest pain*) were dichotomous versions of multivalued variables (ie, *primary symptom*, which has 10 possible values). In such cases, only the multivalued variable was used. All together, the decision trees were generated from the remaining 52 variables.

The variables have a variety of simple and complex relationships among them, providing a challenging and realistic domain for developing decision-aids. They include dichotomous variables (eg., presence of chest pain), multivalued variables (eg., primary symptom), and continuous variables (eg., heart rate). Many variables are further characterizations of other variables, such as the location, type, and degree of chest pain. Some of the variables are closely related, but not so clearly. For example, there were two data collection forms filled on each patient, one from the emergency-room medical record and one through an interview conducted by a research assistant hired for that purpose. Both of these worksheets contain information about chest pain. The emergency-room worksheet asks if the patient presented with chest pain while the interviewer asked if the patient had any feeling of discomfort, pressure, or pain in the chest. Logically, all for whom the first was true should be included in the second, but the realities of multiple sources and fallible historians means that there are a few exceptions. The final diagnosis of the patient is a categorical variable with 25 ordered values. That is, the patient received the classification of the first value in the diagnosis list that was true. The first eight of these values are four severities of myocardial infarction (the Killip classes) and four severities of angina pectoris (New York Heart Association classes). The rest of the values include other types of heart disease, other diseases, and other possible causes for the presenting complaints. For purposes of this study and the development of the LR instrument, the myocardial infarction and unstable angina values are considered to be acute ischemia and the rest of the values are not.

4 Comparison of Methods

Since the yes/no classification output provided by a decision tree is different from the probability of the class provided by an LR equation, the first problem is how to compare the two methodologies. It is possible to transform either of the outputs into the other, but incurring some loss of information. Since each transformation may favor one or the other methodology, both transformations must be considered.

The LR equation can be transformed into a tree by considering some probability to be the threshold for a classification of acute ischemia. That is, any combination of values of the seven variables with a computed probability less than the threshold is classified as not having acute ischemia and those greater are classified as having acute ischemia. If the decision were simply which alternative is most likely, the appropriate threshold would be approximately 0.5 (depending on the prevalence and the probability distributions). However, more risk to the patient is incurred by missing the diagnosis than by making it falsely, so a threshold less than 0.5 is more appropriate. Even adjusting the threshold, this transformation forfeits some of the power of the LR equation because it lumps the patient with a 1% likelihood of acute ischemia and the patient with a nearly 50% likelihood as not having ischemia. Similarly, the patient with a 51% likelihood of ischemia is classified with the patient in whom it is virtually certain. One could argue that that is not a problem, because some decision has to be made in each case, but the information from this decision tool will not be used alone. It will be combined with physician judgement to decide among many options, including various levels of care and further testing. Therefore, certainty of the diagnosis does make a difference. Indeed, the LR equation was originally developed with the objective of helping physicians improve their specificity, rather than some other objective such as minimizing the error rate.

The decision tree can be transformed to provide a probability instead of a classification by using the actual distribution of the patients in each leaf of the tree as the probability of ischemia, rather than assigning the leaf entirely to the most frequently occurring category. If there are a significant number of patients at each leaf node, this will give a reasonable approximation of the probability. The problem is that with a large number of leaf nodes, there will be some nodes with a small number of patients and correspondingly less accuracy in the probability estimates. Furthermore, the frequency count does not take into account any information there might be in the statistical relationships with variables higher in the tree. That is, if the variable were independent of the variables higher in the tree, a better estimate of the probability could be obtained from the whole data set than from the subset at the current node. Since ID3 makes no independence assumption, it throws away this potential source of information. This issue has been addressed in the literature[17, 18] and we have adopted the solution suggested by Quinlan. That is, the estimated probability is adjusted by the probability in the context of the leaf node, where the context is the set of patients that differ from the requirements of the leaf node by at most one variable value. If there are n' patients in the context, i' of which have acute ischemia, n patients at the leaf, i of which have acute ischemia, the probability of acute ischemia is estimated to be

$$\frac{i(n'-i')+i'}{n(n'-i')+n'}$$

4.1 Default Decision Tree

The first comparison is the tree generated by ID3 given the variables in the learning set using the default parameter settings compared to the classifications implied by the LR equation probabilities. The decision tree is quite large, having 312 leaf nodes. The top two levels of the decision tree are shown in figure 1 with the numbers of patients in the learning set with and without ischemia in parentheses. In contrast the seven variables used in the LR equation with their beta exponents (for true = 1 and false = 2) are shown in figure 2. Both of these sets of variables include the *ST change* and *Chest pain in 24 hours*. The other four variables in the LR equation appear further down in the tree among the 43 variables used. Two of the four remaining variables appearing at the second level of the tree are closely related to variables in the LR equation. Obviously, the tree makes no attempt to minimize the number of variables used. For example, where *chest pain in the ER* was used, *chest pain in 24 hours* was the second choice. Since the LR equation is using dichotomous variables (although that is not required), the *ST change* variable can be used more than once with different divisions of the values. The two *ST change* variables in the LR equation

Key: variable = value (I:32 NI:18) means 32 had acute ischemia and the remaining 18 did not.

ST change on ECG = normal (I:395 NI:1646)
    Nitroglycerin stops pain = yes (I:194 NI:234)
    Nitroglycerin stops pain = no (I:201 NI:1412)
ST change on ECG = down 2mm (I:117 NI:49)
    Dizzy in the ER = yes (I:12 NI:16)
    Dizzy in the ER = no (I:104 NI:33)
ST change on ECG = down 1mm (I:150 NI:148)
    Chest pain in 24 hours = yes (I:136 NI:93)
    Chest pain in 24 hours = no (I:14 NI:55)
ST change on ECG = down 0.5mm (I:80 NI:120)
    Nitroglycerin stops pain = yes (I:38 NI:12)
    Nitroglycerin stops pain = no (I:42 NI:106)
ST change on ECG = flat (I:63 NI:95)
    Nitroglycerin stops pain = yes (I:33 NI:11)
    Nitroglycerin stops pain = no (I:30 NI:84)
ST change on ECG = up 1mm (I:209 NI:108)
    Age $<$ 87.5 (I:208 NI:101)
    Age $>$ 87.5 (I:1 NI:7)
ST change on ECG = up 2mm (I:238 NI:35)
    Chest pain in the ER = yes (I:205 NI:16)
    Chest pain in the ER = no (I:32 NI:18)

Figure 1: Top Levels of Default Decision Tree

LR variable	value	β coefficient
T waves on ECG	normal or flat	1.1278
Chest pain in 24 hours	true	0.9988
ST change on ECG	normal	0.8321
ST change on ECG	normal or flat	0.7682
Primary symptom	chest pain	0.7145
History of nitroglycerin use	true	0.5091
History of myocardial infarction	true	0.4187

Figure 2: LR Regression Coefficients

		actual class			
		learning set		test set	
		ischemia	no ischemia	ischemia	no ischemia
ID3 class	ischemia	1,092	111	455	304
	no ischemia	160	2,090	258	1,305
	error rate		0.0776		0.2470
LR class	ischemia	773	295	518	183
	no ischemia	479	1,906	193	1,421
	error rate		0.2242		0.1624

Figure 3: Comparison of the Error Rates Between LR and ID3

have the greatest influence on the probability, as would be expected from the decision tree. Thus, the choice of variables between the two methods is closely related but not identical.

The comparison of the correct and incorrect classifications for the two strategies (using a threshold of 0.5) is shown in figure 3. Note that in the test set, only 2,315 of the 2,320 cases are classified by LR. The remaining five do not have electrocardiograms. LR has no mechanism for classifying a patient when there is missing data, while ID3 picks the most likely classification given the proportions of cases with the different possible values of the missing data.

The classification threshold of 0.5 does not optimize patient benefit since missing acute ischemia is much worse than falsely assigning acute ischemia. Thus, it is important to consider the effect of changing the cutoff for classification. Figure 4 is a plot of the cutoff probability versus the error rate. Included in the figure is the performance of both the decision tree and the LR equation on both the learning set and the test set. If the decision tree were unpruned and there were no cases with identical variable values and different diagnoses, the ID3 curve for the learning set would have an error rate of zero. With a pruned tree, the curve gives an approximate limit on how well a tree of that size can classify the cases (at least when variables are chosen one at a time). That the performance of ID3 on the learning set is considerably different from the performance on the test set indicates the tree is overspecified. That is, some of the branches in the tree are tracking the statistical random variation in the learning set, rather than the underlying properties of the population. The error rate curves for ID3 are essentially flat from a threshold of 0.25 to 0.8. This too is an indication that the tree is overspecified. Using a large number of nodes, ID3 has found small partitions of the learning set in which the elements are nearly all ischemic or nearly all not ischemic. Indeed, in the learning set there are only 359 of the 3,453 patients whose estimated probability of ischemia is between 0.25 and 0.8. In the test set 278 of the 2,320 cases fall in this range.

Comparing the error rate of the LR classifications to the decision tree classifications, it is clear that the LR equation performs better on the test set over the threshold range from 0.28 to 0.77. Over the range from 0.37 to 0.67 the improvement in error rate is about 25%. It is curious that the LR equation actually works better on the test set than the learning set. Since these two data sets were collected in consecutive years, it is possible that the patient population changed somewhat, but there is no way to tell for sure. The incidence of acute ischemia did drop from 36% to 31%, but that is a rather small change. Even comparing the LR equation on the learning set to the decision tree on the test set, the LR equation did better with thresholds between 0.4 and 0.55, although only slightly better.

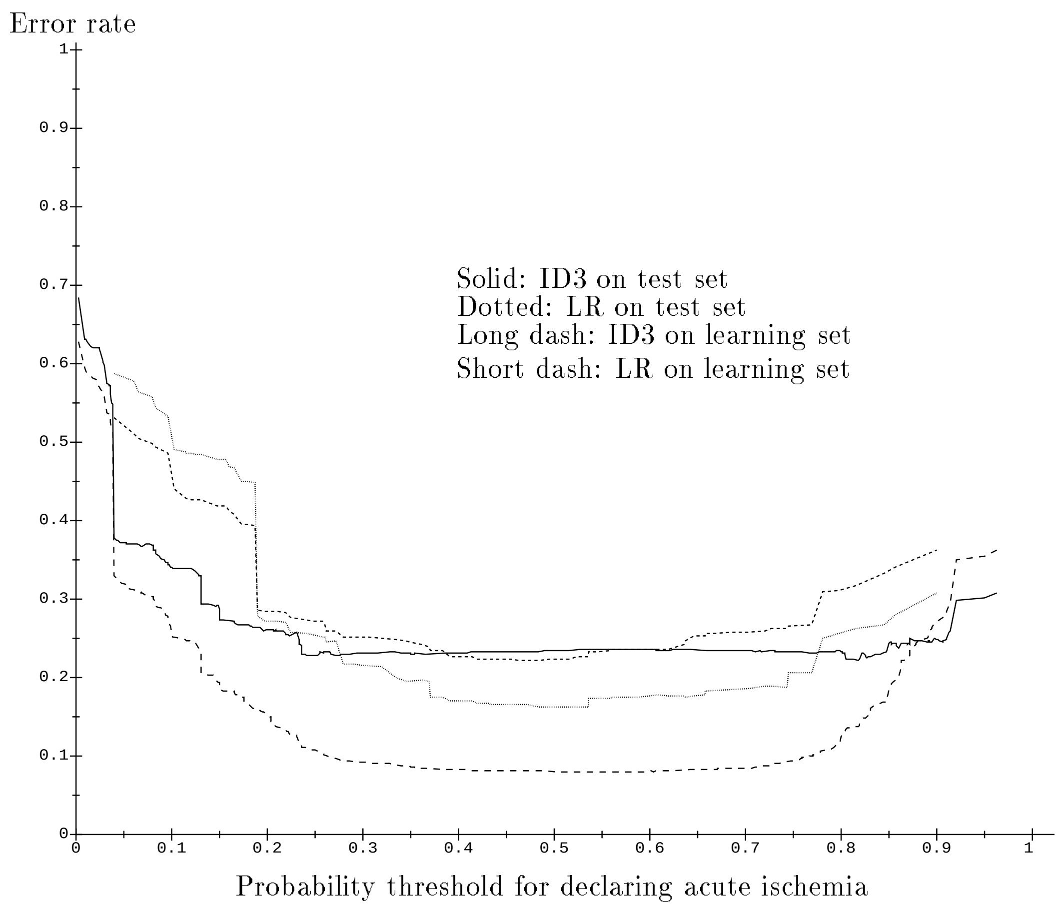

Figure 4: Error Rate with Varying Cutoff Probability for Default Tree

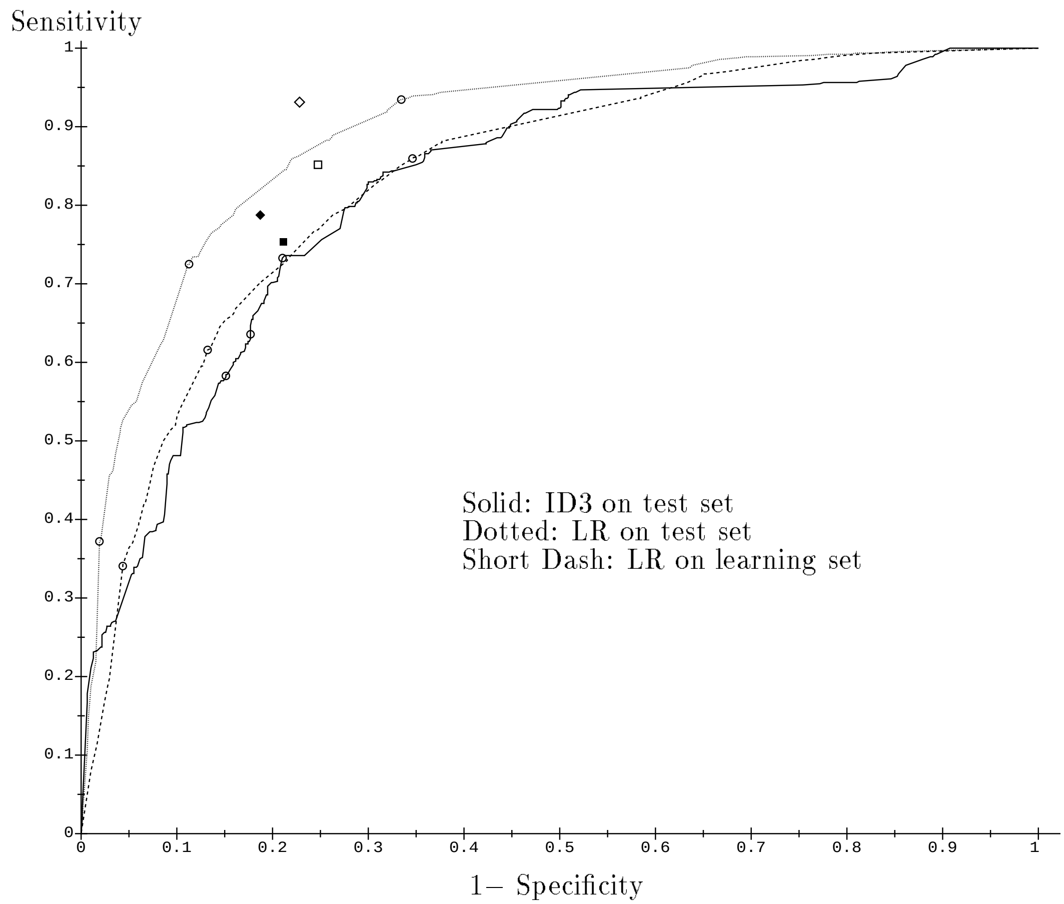

Figure 5: Receiver Operating Curve for Default Tree

Another way to look at these data, that is less dependent on the frequency of ischemia in the population, is to consider the sensitivity and specificity of the tree at various probability levels. The usual method for doing this is to plot the sensitivity versus 1− specificity as an ROC (receiver operating characteristic) curve. This is done in figure 5. The performance of LR on both the test set and the learning set is graphed, showing the improvement in performance on the test set. The area under the ROC curve for LR on the test set is 0.89 while that under the ID3 curve is 0.82. The difference between these two areas is significant well beyond the .0001 level using the Hanley-McNeil method with correlations computed using the Kendall tau[19].

Since the other point of comparison is the performance of the physicians in the emergency room, their performance is indicated by the markers on the graph. The squares are the performance on the learning set and the diamonds are the performance on the test set. It is clear that the physicians also performed better on test set than on the learning set. Thus, the difference between sets at least involves either differences in the population or differences in physician behavior. One of the categories of diagnosis used by the physicians is *ischemic heart disease*, without indicating whether it is acute. The open markers include this diagnosis as acute ischemia while the closed markers exclude it. Comparing the open and closed markers gives an indication of what an “ROC curve” for the physicians might look like.

tree	data	ischemia	non ischemia	difference
default ID3	test	0.61	0.22	0.39
LR	test	0.62	0.24	0.38
default ID3	learn	0.77	0.13	0.64
LR	learn	0.57	0.25	0.32

Figure 6: Average Probabilities Assigned to Cases

On the curves the points corresponding to thresholds of 0.25, 0.5, and 0.75 are indicated with circles. Comparing the LR curves to the ID3 curve, the categorical nature of the decision tree is evident from the angularity of the curve and how close the threshold points are. It is clear that when both sensitivity and specificity are considered, the LR equation performs better.

The ideal for a tree is to classify all of the test cases correctly. When that is not possible, the objective is to provide as much separation as possible between the two classifications of cases. One way to test this is to compare the average probability of ischemia given to actual ischemic cases to that given to the non-ischemic cases. Figure 6 shows the average probabilities given to cases that were ischemic or non-ischemic in the test set and in the learning set by the two methods. By this measure, the decision tree does better, but only very slightly. The primary difference is that the "overconfidence" of the decision tree compensates for the higher error rate. That is, it is right often enough that the extreme probabilities assigned by leaves with small numbers of cases actually achieves a greater separation of average probabilities. The performance of ID3 on the learning set again demonstrates the amount of separation of ischemic from non-ischemic cases that can be achieved with foreknowledge using a tree of a given size. The performance of the LR equation on the learning set shows that the equation actually did better on the test set than on the data from which it was derived.

4.2 Improving the Decision Tree

Since it is clear from this analysis that the decision tree is overspecified, we reconsidered the options available with the ID3 algorithm at the risk of "learning from the test set". In the following, we set the test set aside and proceeded just using the learning set to motivate and evaluate adjustments to the ID3 parameters and improvements to the selection of variables.

One strategy developed to generate trees that are better justified is *windowing*. Windowing generates the initial tree from a subset of the data and uses the rest of the data to modify the tree as necessary to account for any statistical differences between that data and the initial subset. If several such trees are created using different initial subsets and the best final tree is chosen, the claim is that the tree will better represent the true statistical properties of the data. We tried this strategy by generating five trees using 2/3 of the total learning set as the learning set for the experiment and the remaining 1/3 as the test set. Each tree was started with windows containing 10% of the data. The error rates for the five trees are given in figure 7. These error rates are no better than the error rate with the tree generated without windowing. These conclusions concur with those of Wirth and Catlett[20]. Thus, we have chosen not to use windowing for generating trees.

If one examines the default tree, it is clear that there are many places where branches divide the cases into a very small set versus the rest. For example, at the third level in the tree, the following

tree	1	2	3	4	5
learning errors	0.0750	0.0770	0.0773	0.0785	0.0773
test errors	0.2453	0.2513	0.2379	0.2332	0.2534

Figure 7: Error Rates Using Windowing

happens:

ST change on ECG = normal
    Nitroglycerin stops pain = yes
        age $<$ 36.5 : ischemia (3 of 3)
        age $>$ 36.5 (the remaining 425)
            ...

While it may well be that a person under 37 complaining of chest pain in the emergency room for whom nitroglycerin stops the pain is a strong candidate for acute ischemia, three is too small a sample to produce a node. To overcome this problem, we modified the tree generation to only consider nodes with six or more cases in two or more branches.

Furthermore, the tree generated is very large, having a maximum depth of 25. The method used for pruning the tree is to compare the number of errors introduced by eliminating the branch, to the estimated error rate in the subtree using the continuity correction for the binomial distribution plus the standard error of the estimate. It is possible to make the pruning more optimistic or pessimistic by adjusting the number of standard errors added into this calculation. To estimate an appropriate number for this parameter, we compared the effects of different amounts of pruning on various trees generated from different partitions of the original learning set into 2/3 as the learning set and the remaining 1/3 as the test set.

Figure 8 graphs three of these tests using pruning ranging from none to eight times the standard error. When pruning levels reach about five, the trees have about ten variables in them and about 30 leaves, although it varied considerably. The vertical axis is the number of errors classifying the test set. The trends show that the error rates initially decrease with increased pruning but then start to fluctuate. Since the error rate is consistently fairly low when the pruning parameter is set to 2.0 standard errors and the tree size is comparable to the tree implied by the LR equation, we chose that value for generating the tree from the entire learning set.

4.3 The Improved Decision Tree From ID3

With a pruning parameter of 2.0 and minimum leaf size of six in at least two branches of a variable, the tree for the following comparisons was generated. It has 128 nodes, 88 of which are leaf nodes and 70 of these have items corresponding to them in the learning set. The maximum depth of the tree is 11, and it uses 19 different variables. This compares to the tree generated by the 128 possible value combinations of the seven LR variables with 85 leaves having corresponding items. The correct and incorrect classifications for these two trees are shown in figure 9.

This decision tree can be compared to the default tree and to the LR equation by considering what would happen if the threshold for classification were changed. Figure 10 is a plot of the threshold probability versus the error rate. This time the performance of ID3 on the test set is

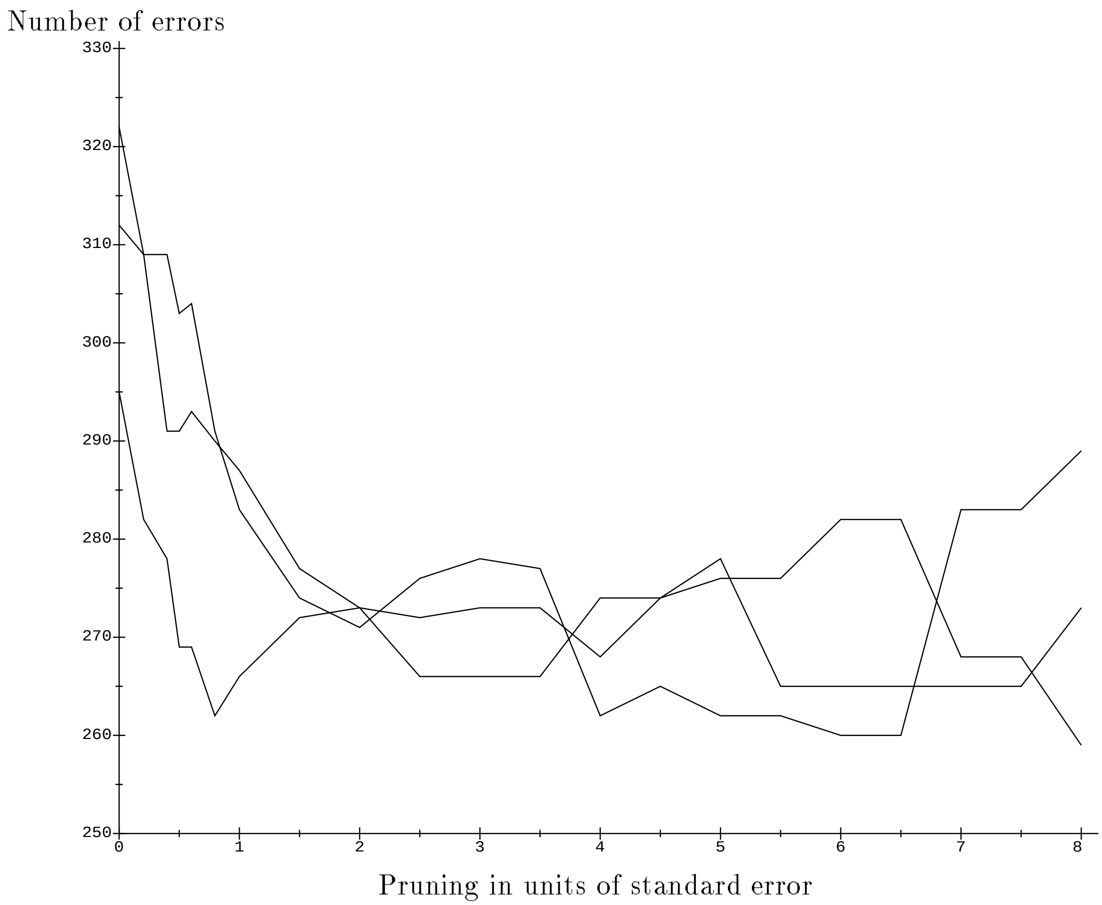

Figure 8: Error Rates with Different Amounts of Pruning

		actual class			
		learning set		test set	
		ischemia	no ischemia	ischemia	no ischemia
ID3 class	ischemia	913	251	471	219
	no ischemia	339	1,950	242	1,388
	error rate		0.1710		0.1987
LR class	ischemia	773	295	518	183
	no ischemia	479	1,906	193	1,421
	error rate		0.2242		0.1624

Figure 9: Error Rates Between LR and ID3 with More Pruning

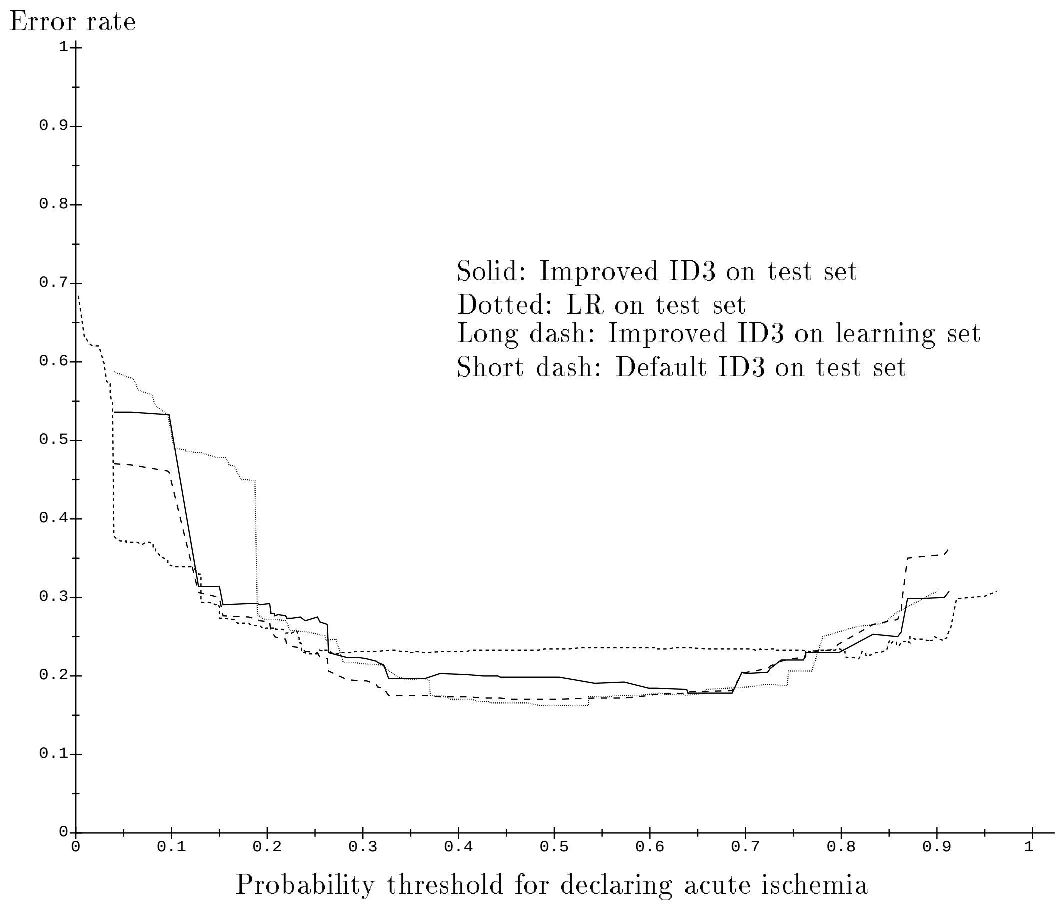

Figure 10: Error Rate with Varying Cutoff Probability in Improved Tree

much closer to its performance on the learning set, although it still performs better on the learning set. The performance of ID3 on the test set is improved over the default tree and the flat portion of the curve is narrower, from 0.33 to 0.69. There are now 272 cases in the learning set with probabilities in this range and 197 in the test set, but in the same range the LR equation has 977 and 640 cases, respectively. When the cutoff is between 0.35 and 0.65, the LR equation still performs better on the test data than the improved decision tree does. Outside of that range, the error curves overlap considerably.

Again, the ROC curves can be plotted for the improved tree, along with the previous curves. This is done in figure 11. The ROC curve for the improved decision tree lies almost completely outside of the ROC curve for the default decision tree and still completely inside the ROC curve for the LR equation. Thus, no matter which operating point on the curve is chosen, the LR equation has better sensitivity and specificity. The LR curve has an area of 0.89 while the improved ID3 curve has an area of 0.86 and the difference in area is still significant at the .0001 level. The improvement in the area under the ID3 curve is significant at the .005 level. The points corresponding to thresholds of 0.25, 0.5, and 0.75 are also plotted on this graph. With the improved decision tree, these thresholds correspond to sensitivities and specificities that are considerably different.

If we consider the degree to which the improved decision tree separates the ischemic from non-

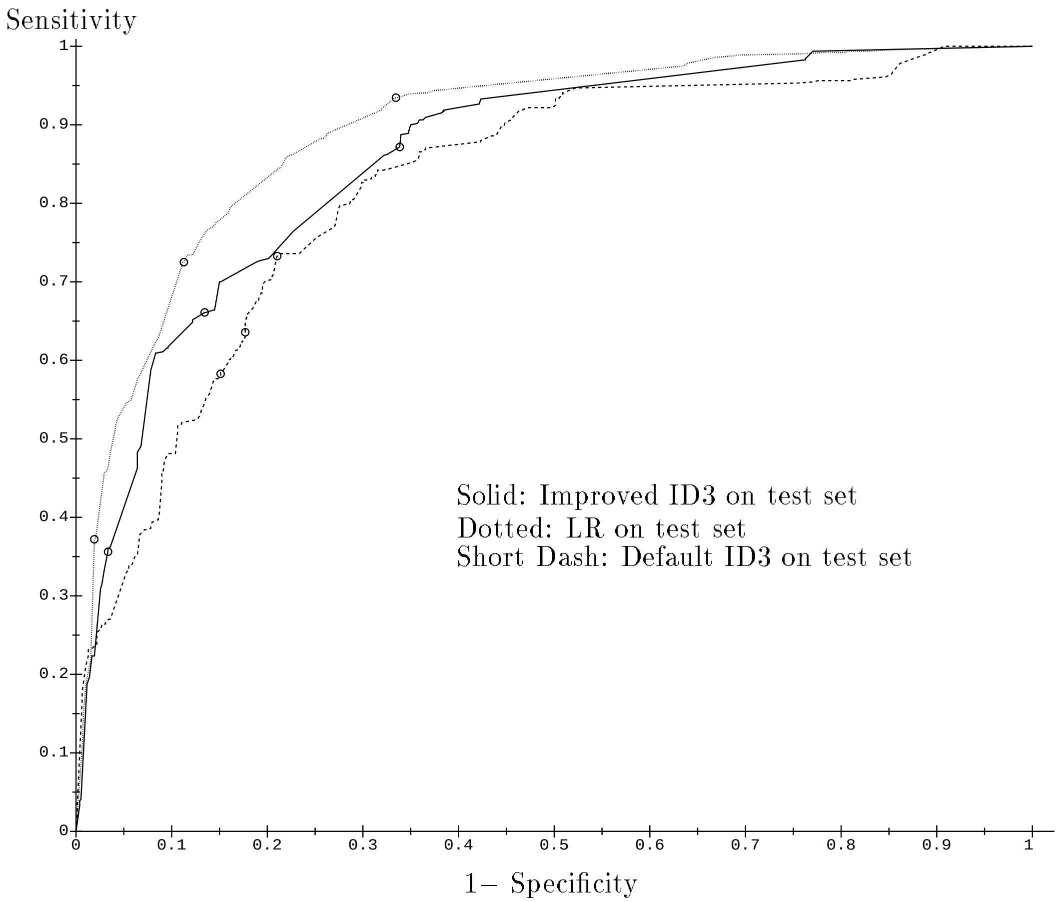

Figure 11: Receiver Operating Curve for Improved Tree

tree	data	ischemia	non ischemia	difference
improved ID3	test	0.60	0.22	0.38
default ID3	test	0.61	0.22	0.39
LR	test	0.62	0.24	0.38
improved ID3	learn	0.63	0.21	0.42

Figure 12: Average Probabilities Assigned to Cases by Improved Decision Tree

ischemic cases (in figure 12), there is virtually no change from the default tree and it is essentially the same as the LR equation. The only difference is that the performance of the improved tree on the learning set is more in line with the performance on the test set.

4.4 Decision Tree with Clinical Considerations

The improved decision tree generated in the previous section reduced the problem of overspecification and as a result, improved the performance of the tree. However, it was still developed using all of the variables that were ever considered for the LR equation without consideration for clinical relevance. When the LR equation was developed, the clinical relevance and reproducibility of the variables was a primary consideration and many of the variables were eliminated on that basis alone[2]. Since no such considerations were made in producing the decision tree, it contains a number of closely related variables such as the report of chest pain in the emergency room and the indication of chest pain recorded in a later interview. It also contains variables such as "upset" that would be difficult to reproduce.

To make a tree that is more relevant clinically, a limited set of variables was chosen from which to generate it. We started with the variables that were actually used in the previous decision tree plus any that had significant information gain ratios across the whole learning set. Thus, all variables with power to discriminate over the whole set were considered and additional variables that had emerged at lower levels were also considered. This list was then modified by using only variables collected in the emergency room versus corresponding variables collected later and selecting one of the two nitroglycerin variables. From this list upset, dizzy, and palpitations were eliminated because of the subjective nature of the finding, race was eliminated because of the small fraction of non-whites, and history of bronchitis was eliminated on grounds of clinical relevancy. This left a list of 15 variables from which the tree was generated using the same procedure as in the previous section.

The resulting tree has 103 nodes, 72 of which are leaf nodes and 62 of these have items corresponding to them in the learning set. The maximum depth of the tree is nine, and it uses 11 of the 15 variables available to it. Of the 11 variables used, two were not used in the previous tree (affecting 7% of the classifications in the learning set) and two others did not have significance at the top level (affecting 11% of the classifications).

The leaf nodes without items corresponding to them in the learning set represent values for the three multivalued variables. Since there are other values for these variables that have only a few cases, we further trimmed the tree by comparing all nodes corresponding to values of multivalued variables to the nodes immediately above them. If by the chi-square test, the difference in frequency of ischemia was not significant, the probability implied by the parent node was used. In this way, another 19 leaves were eliminated.

This clinical decision tree can be compared to the improved decision tree and to the LR equation by considering the error rates across the range of thresholds for classification in figure 13. Even though the selection of variables was restricted, the performance of this tree is slightly better than the previous tree across the range 0.35 to 0.68. The performance on the test set is essentially the same as the performance on the learning set, indicating that there is no problem of overspecification. The LR equation still maintains an edge on the error rate between 0.37 and 0.6.

The ROC curve for the clinical decision tree overlaps considerably the ROC curve for the improved decision tree and still completely inside the ROC curve for the LR equation, as can be

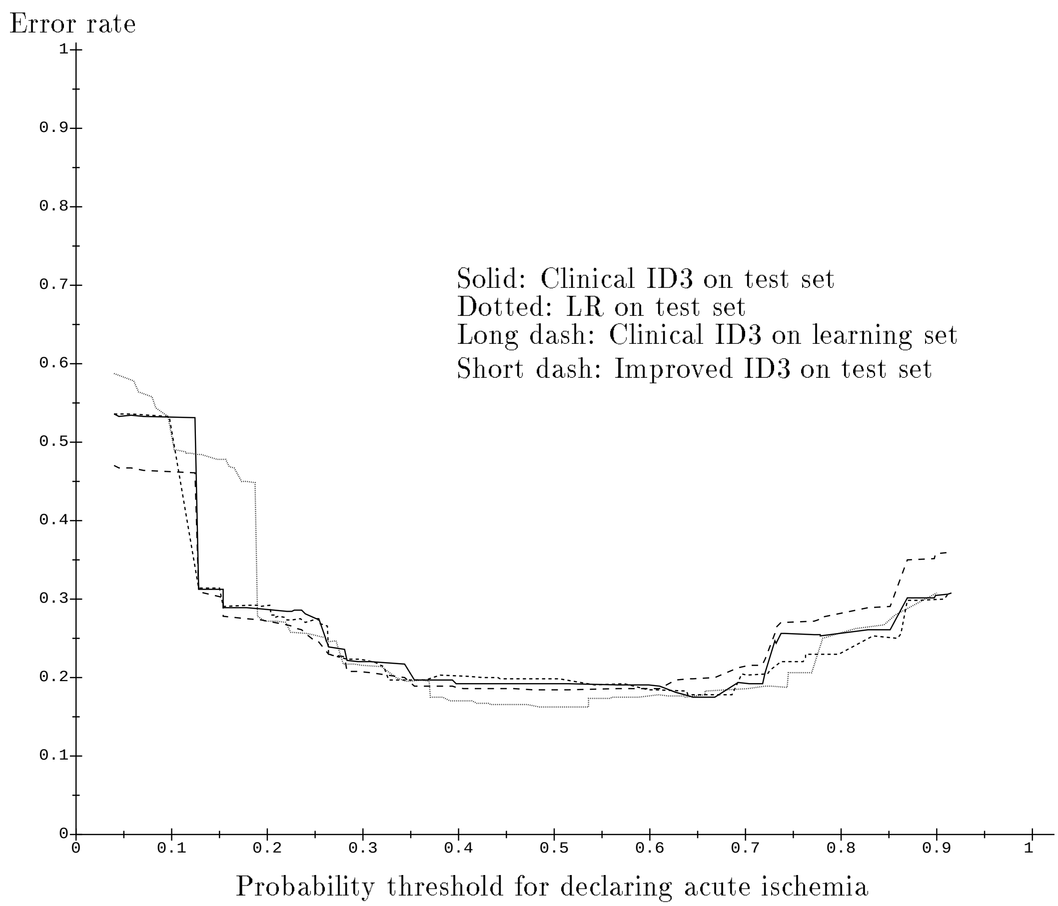

Figure 13: Error Rate with Varying Cutoff Probability in Clinically Based Tree

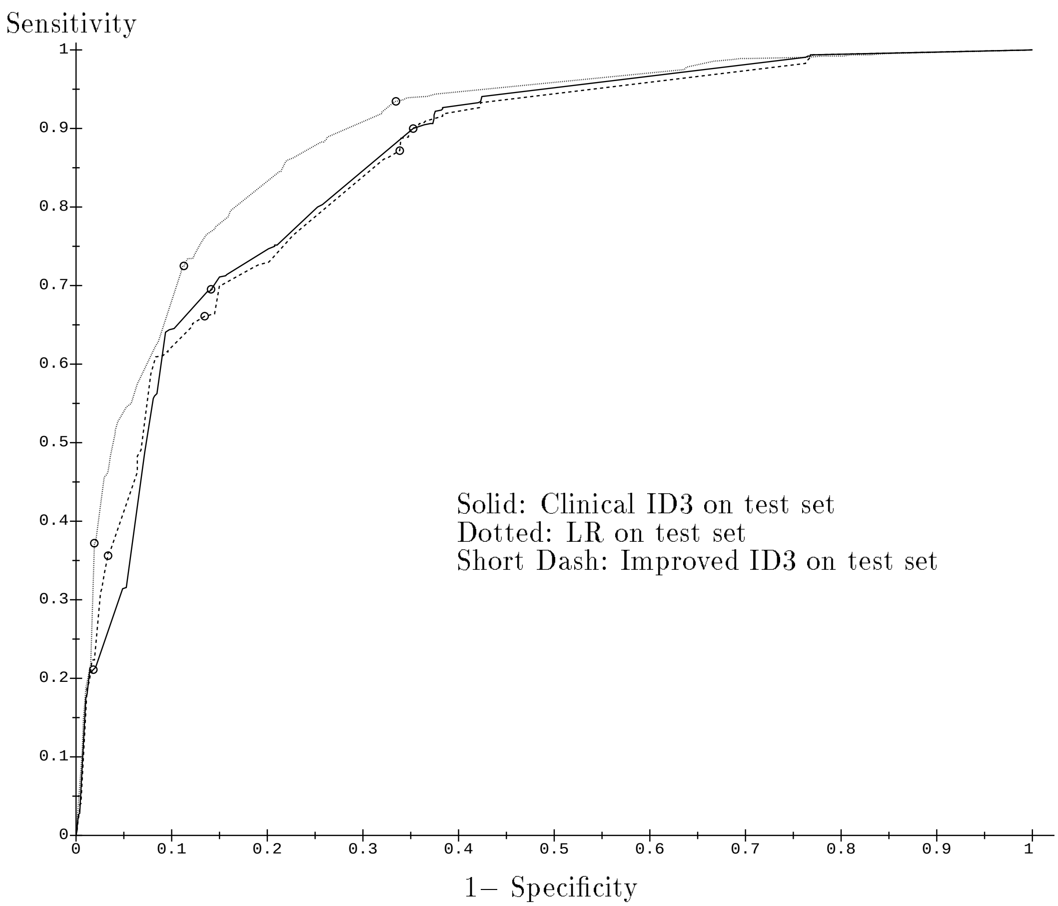

Figure 14: Receiver Operating Curve for Clinically Based Tree

seen in figure 14. Thus, no matter which operating point on the curve is chosen, the LR equation has better sensitivity and specificity. The LR curve has an area of 0.89 while the clinical decision curve has an area of 0.86, the same as that for the improved tree. The difference in areas with respect to the LR curve remains significant at the .0001 level, while there is no significant difference in areas between the two ID3 curves.

5 Discussion

The comparison of tools developed using logistic regression and using decision-tree induction presents an interesting contrast of statistical assumptions. For LR, it is assumed that the influence of a variable on the outcome is uniform across all patients unless specific interactions with other variables are included. For ID3 and decision-tree formalisms in general, for each subset of the patients as defined by the particular values of the variables higher in the tree it is assumed that the effect of a variable in the subset is unrelated to the effect of the variable in other subsets of patients. Since these are polar assumptions and the truth probably is somewhere in between, it makes sense to ask which is closer to the truth. The answer from this examination of a data base of over 5,000 patients is that both methodologies are capable of performance fairly close to that of the physicians who treated the patients, although the LR equation definitely performed better.

The default decision tree was much larger than was justified by the data. This can be attributed to a pruning strategy that does not capture the statistical nature of the data appropriately. Other classification tree programs use other strategies for pruning which may be more appropriate and should be investigated. As was clear from the comparison made in section 4.3, optimizing the parameters of the pruning algorithm on the learning set improved the performance of the decision tree and eliminated any obvious overspecification. Indeed, the overspecification of the default tree is responsible for about 56% of the difference in error rate between LR and ID3. The underlying premise of the information theory approach to favor the variables that achieve the most extreme concentrations of the dependent variable values is still evident in the relatively small number of cases that are in the middle of the probability range.

The final decision tree, generated with clinical considerations in mind, demonstrates that there is very little difference between the performance of a tree generated with all of the variables available and one using a greatly restricted set. One could argue either that restricting the set would force suboptimal choices or that using using clinical knowledge to restrict the set would avoid spurious correlations and improve the choices. The overall effect seems to be negligible. Even the final tree (reproduced in the appendix) still has what are likely to be spurious correlations. For example at the end of the tree four levels down, Q waves in the anterior leads would indicate no ischemia while no Q waves and a heart rate of less than 69 would indicate ischemia, which is counter intuitive. Such questionable leaves involve only small numbers of cases, indicating that there is still room for improvement in deciding what relationships are legitimate.

Since there is considerable overlap between the variables used in the LR equation and the decision tree, it might be suspected that the two procedures are picking the same cases. However, in comparing the probabilities generated for the cases, 23% of the learning cases and 20% of the test cases have probabilities that differ by more than 0.2. Also, the Kendall tau correlation of the assigned probabilities is only 0.50. Both of these measures indicate that there may be features of the domain that each method is missing and therefore potential for finding a method that would improve on both.

Another difference between the methodologies is the way in which the probabilities are calculated. The LR equation is derived from the entire learning set, while the ID3 probability is estimated from the leaf subset. With a sufficiently large number of cases in each leaf subset, the fraction of ischemia cases in the subset is the correct probability. However, with a limited data base, the probabilities are approximate, although the correction for small subsets using the immediate context improves the estimates considerably. We can evaluate the LR estimate of the probability by comparing the estimated probabilities. The number of cases where the LR estimate is closer to the fraction in the test set is 958, while the number in which the fraction in the learning set is closer to the fraction in the test set is 1,352. However, if these numbers are weighted by the size of the differences, the sum of the differences in which the LR estimate is closer is 74 while the sum in which the learning set fraction is closer is 65. Thus, the ID3 style probability is closer more often, but the size of its misses more than makes up for that. The probabilities might be improved either by including combinations of values as variables in the LR equation or by including more context in the ID3 probability estimations. The challenge is to develop criteria for determining the appropriate compromises between the strategies.

The final question is whether either of these methods have extracted all of the clinical relationships that are supported by the data. Given the experience of many researchers attempting to detect acute ischemia or myocardial infarction from information in the emergency room, there seems to be a limit to the reliability of the results and given how similarly these two methods perform, we may be near that limit. However, there are always new strategies to consider. Two that may hold promise are neural networks[21] and multivariate adaptive regression splines[22]. Both of these methods have the potential to represent more complex relationships, but with more flexibility in the model there is more potential for overspecification.

References

[1] Pozen, M. W., D'Agostino, R. B., et al, "The Usefulness of a Predictive Instrument to Reduce Inappropriate Admissions to the Coronary Care Unit," *Annals of Internal Medicine* 92:238-242, 1980.

[2] Pozen, M. W., D'Agostino, R. B., et al, "A Predictive Instrument to Improve Coronary-Care-Unit Admission Practices in Acute Ischemic Heart Disease," *New England Journal of Medicine* 310:1273-1278, 1984.

[3] Selker, Harry P., Griffith, John L., and D'Agostino, Ralph B., "A Tool for Judging Coronary Care Unit Admission Appropriateness Valid for Both Real-Time and Retrospective Use: A Time-Insensitive Predictive Instrument (TIPI) for Acute Cardiac Ischemia," forthcoming.

[4] Goldman, L., Weinberg, M., et al, "A Computer-Derived Protocol to Aid in the Diagnosis of Emergency Room Patients with Acute Chest Pain," *New England Journal of Medicine* 307:588-596, 1982.

[5] Goldman, L., Cook, E. F., et al, "A Computer Protocol to Predict Myocardial Infarction in Emergency Department Patients with Chest Pain," *New England Journal of Medicine* 318:797-803, 1988.

[6] Breiman, L., Friedman, J. H. et al, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.

[7] Mingers, J., "Rule Induction with Statistical Data — A Comparison with Multiple Regression," *J. Operational Research Society* 38:347-351, 1987.

[8] Segal, M. R. and Bloch, D. A., "A Comparison of Estimated Proportional Hazards Models and Regression Trees," *Statistics in Medicine* 8:539-550, 1989.

[9] Harrell Jr., F. E., Lee, K. L. et al, "Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions," *Cancer Treatment Reports* 69:1071-1077, 1985.

[10] Kinney, E. L. and Murphy, D. D., "Comparison of the ID3 Algorithm versus Discriminant Analysis for Performing Feature Selection," *Computers and Biomedical Research* 20:467-476, 1987.

[11] Gilpin, A. E., Ohlsen, R. A., "Predicting 1-Year Outcome following Acute Myocardial Infarction: Physicians versus Computers," *Computers and Biomedical Research* 23:46-63, 1990.

[12] Michalski, R. S., "AQVAL/1 – Computer Implementation of a Variable Valued Logic System VL1 and Examples of its Application to Pattern Recognition," *Proceedings of the First International Joint Conference on Pattern Recognition*, Washington, DC, pp 3-17, 1973.

[13] Quinlan, J. R., "Induction of Decision Trees," *Machine Learning* 1:81-106, 1986.

[14] Mingers, J., "An Empirical Comparison of Selection Measures for Decision-Tree Induction," *Machine Learning* 3:319-342, 1989.

[15] Quinlan, J. R. and Rivest, R. L., "Inferring Decision Trees Using the Minimum Description Length Principle," *Information and Computation* 80:227-248, 1989.

[16] Quinlan, J. R., "Simplifying Decision Trees," *International Journal of Man Machine Studies* 27:221-234, 1987.

[17] Quinlan, J. R., "Improved Estimates for the Accuracy of Small Disjuncts," *Machine Learning* 6:93-98, 1991.

[18] Holte, R. C., Acker, L. E., and Porter, B. W., "Concept Learning and the Problem of Small Disjuncts," *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, pp 813-818, 1989.

[19] Hanley, J. A., and McNeil, B. J., "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases," *Radiology* 148:839-843, 1983.

[20] Wirth, J. and Catlett, J., "Experiments on the Costs and Benefits of Windowing in ID3," *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, MI, pp87-99, 1988.

[21] Baxt, W. G., "Use of an Artificial Neural Network for the Diagnosis of Myocardial Infarction," *Annals of Internal Medicine* 115:843-848, 1991.

[22] Friedman, J. H., "Multivariate Adaptive Regression Splines," *The Annals of Statistics* 19:1-141, 1991.

A Clinically Motivated Decision Tree Used for Comparison

The following is the final decision tree generated from variables selected on clinical grounds.

- STCHANGE = $\uparrow$ 2 (I: 238 NI: 35)
- STCHANGE = NORMAL (I: 395 NI: 1646)
  - NCPNITRO = YES (I: 194 NI: 234)
    - CHPAINER = NO (I: 7 NI: 22)
    - CHPAINER = YES (I: 187 NI: 212)
      - SYMPTOM1 = pain in arms, neck, or shoulders (I: 11 NI: 4)
      - SYMPTOM1 = shortness of breath (I: 9 NI: 33)
      - SYMPTOM1 = pain in stomach (I: 7 NI: 10)
        - SEX = MALE (I: 1 NI: 8)
        - SEX = FEMALE (I: 6 NI: 2)
      - SYMPTOM1 = pressure, pain, discomfort in chest (I: 147 NI: 152)
        - SEX = FEMALE (I: 45 NI: 81)
        - SEX = MALE (I: 102 NI: 71)
          - AGE $>$ 81.5 (I: 1 NI: 6)
          - AGE $<$ 81.5 (I: 101 NI: 65)
            - AGE $<$ 45.5 (I: 6 NI: 9)
            - AGE $>$ 45.5 (I: 95 NI: 56)
  - NCPNITRO = NO (I: 201 NI: 1412)
    - CHPAINER = NO (I: 25 NI: 604)
    - CHPAINER = YES (I: 176 NI: 807)
      - TWAVES = NORMAL (I: 76 NI: 462)
      - TWAVES = $\downarrow$ 1 (I: 63 NI: 62)
        - SEX = FEMALE (I: 20 NI: 38)
        - SEX = MALE (I: 43 NI: 24)
          - AGE $<$ 73.5 (I: 39 NI: 14)
          - AGE $>$ 73.5 (I: 4 NI: 10)
- STCHANGE = $\downarrow$ 2 (I: 117 NI: 49)
  - NCPNITRO = YES (I: 57 NI: 11)
  - NCPNITRO = NO (I: 60 NI: 38)
    - SYSBP $>$ 202.0 (I: 8 NI: 0)
    - SYSBP $<$ 202.0 (I: 51 NI: 37)
      - QWAVES = ASMI (I: 4 NI: 0)
      - QWAVES = NORMAL (I: 37 NI: 28)
        - SYSBP $>$ 178.0 (I: 9 NI: 2)
        - SYSBP $<$ 178.0 (I: 28 NI: 26)
          - AGE $>$ 83.5 (I: 5 NI: 1)
          - AGE $<$ 83.5 (I: 23 NI: 25)
            - PR $<$ 77.0 (I: 2 NI: 7)

```
                        PR > 77.0 (I: 21 NI: 18)
                            PR < 89.0 (I: 8 NI: 0)
                            PR > 89.0 (I: 13 NI: 18)
STCHANGE = ↓ 1 (I: 150 NI: 148)
    SYMPTOM1 = pain in stomach (I: 1 NI: 11)
    SYMPTOM1 = rapid or skipping heartbeats (I: 0 NI: 10)
    SYMPTOM1 = pain in arms, neck, or shoulders (I: 10 NI: 0)
    SYMPTOM1 = shortness of breath (I: 18 NI: 56)
    SYMPTOM1 = fainted, dizziness, or lightheadedness (I: 8 NI: 14)
        AGE > 74.0 (I: 0 NI: 6)
        AGE < 74.0 (I: 8 NI: 8)
            HXMI = YES (I: 6 NI: 1)
            HXMI = NO (I: 2 NI: 7)
    SYMPTOM1 = pressure, pain, discomfort in chest (I: 108 NI: 53)
        PR > 131.0 (I: 0 NI: 6)
        PR < 131.0 (I: 105 NI: 45)
            SYSBP > 197.0 (I: 19 NI: 1)
            SYSBP < 197.0 (I: 86 NI: 43)
                PR < 111.0 (I: 81 NI: 34)
                PR > 111.0 (I: 5 NI: 9)
STCHANGE = ↓ .5 (I: 80 NI: 120)
    NCPNITRO = YES (I: 38 NI: 14)
    NCPNITRO = NO (I: 42 NI: 106)
STCHANGE = FLAT (I: 63 NI: 95)
    NCPNITRO = YES (I: 33 NI: 11)
    NCPNITRO = NO (I: 30 NI: 84)
STCHANGE = ↑ 1 (I: 209 NI: 108)
    AGE > 87.5 (I: 1 NI: 7)
    AGE < 87.5 (I: 208 NI: 101)
        CHPAINER = YES (I: 177 NI: 64)
        CHPAINER = NO (I: 31 NI: 36)
            QWAVES = AMI (I: 0 NI: 4)
            QWAVES = NORMAL (I: 13 NI: 21)
                PR < 69.0 (I: 5 NI: 1)
                PR > 69.0 (I: 8 NI: 19)
```