INTRODUCTION

Malignant melanoma is the most serious of the common skin cancers. The overall number as well as the number per thousand population affected has dramatically increased especially in the fair skinned populations of climates closer to the equator. Direct sun exposure causes the damage to the skin particularly when sustained in the early years of life. The American Cancer Society estimated that about 55,100 new melanomas would be diagnosed in the United States during 2004. The number of new melanomas diagnosed in the United States is increasing. Among white men and women in the US, the incidence of melanoma cases increased sharply at a rate of about 6% per year from 1973 until the early 1980s. Since 1981, however, the rate of increase slowed to slightly less than 3% per year. As with all cancers the earlier the diagnosis and treatment of the lesion the better the prognosis will be. The Breslow thickness, Clark Level and histological type are the three most common pathological measurements that a physician uses to determine the risk of recurrence, or more importantly the risk of metastasis.

Recurrence risk classifications are important for any cancer to limit the use of potentially dangerous adjuvant therapies post local surgical excision to patients who will suffer a recurrence. While the ideal would be to separate these patients accurately with 100% predictability of recurrence, the uncertainties about host response to tumor as well as differences in the tumor biology, constrains us to obtaining the lowest number of false negative predictions while minimizing the number of false positives.

There is considerably more information besides the Breslow thickness, Clark level and histological tumor type available to the treating physician to try and evaluate an individual patient’s prognosis. No definitive method or prediction model is presently available which takes into account the history, physical and other available laboratory data. The major attempts thus far have been to form single and multi-parameter regression models.

In this paper we will use the data from the Duke Melanoma Clinic (DUMC) and the statistical methods of random forests and trees to create a prognostic model in a group of melanoma patients. The goal is to search for a solution to the problem of identifying and separating which patients will most likely have a recurrence from those that will not. The Duke University Melanoma Clinic under the leadership of Dr, Hilliard Seigler, has amassed one of the largest patient databases of melanoma patients in the world. The term "thin melanoma" refers to a subset of the melanoma patients that have lesions less than 1.00 mm. in Breslow thickness. These patients will be the subjects of this paper. The Breslow thickness is attained by the pathologist with a micrometer and is the thickest microscopic measurement of tumor depth found. It is not related to the cross diameter of the typically rounded lesion. The thin melanoma patients have a very low but definite risk of recurrence. Since it has been previously accepted that the Breslow thickness is the most important pathologic measure of risk, in the past patients have been grouped according to this measure. Prior to using the Breslow thickness, the Clark level was the main indicator of the severity of the lesion.

Several papers have been written using this patient collection as well as other patient populations trying to determine which patients are at greatest risk based on the information available. In an article not limited to thin melanomas, Robin Vollmer (2001B) reported that the major factors measuring severity were: thickness, presence of histological ulcer, patient age, sex and body site. He has also reported on the importance of a continuum of thickness levels. Statistical modeling with regression model algorithms applied to patient information have been attempted in order to classify patient risk of recurrence. Using computers, statisticians have been able to examine multiple input variables at one time.

Severity indices such as the hazard score (h.s.), used for risk classification have been formulated. Stadelmann et. al. (1998) reported an equation to measure the 10 year mortality in cutaneous melanoma using just the variable Breslow thickness:

$$\text{h.s.} = 1.9 \cdot [\, 1 - 0.966 \cdot \exp(- 0.2016 \cdot \text{Thickness})] . \tag{1}$$

One hazard score established by multiple regression, using thickness, ulcer, site, age and sex is from Vollmer and Seigler (2001A):

$$\text{h.s.} = 2.28 \cdot \text{Thickness} + 0.472 \cdot \text{Ulcer} + 0.299 \cdot \text{Site} + 0.0622 \cdot \text{Age} - 1.64 \cdot \text{Sex}. \tag{2}$$

A code for the values of the categorical variables is not supplied. They used this formula for the hazard score to predict the 5 yr mortality:

$$\text{5 yr mort} = \frac{1}{1 + \exp(2.59 - 1.42 \cdot \text{h.s.})} . \tag{3}$$

In direct discussions with Dr. Seigler (personal communication, 2005), they learned the predictive value of a linear regression model was best with four variables, two major and two minor parameters although in the hazard score above (2), they used three minor variables. In addition to evaluating recurrence and mortality, other predictive studies have looked at length of time to recur as well as the general survival time. A lack of general consensus in approaching the problem may be best exampled by Gamel, who has written papers on both sides of the argument of whether or not, when predicting patient survival, to use stable or time adjusted algorithms as the length of survival of a patient increases.

Using Duke’s database of 1610 patients, this study attempts to develop complexes of the values of the different input variables collected to try and predict whether or not a newly diagnosed patient will have a recurrence. Other studies such as timing of the recurrence, site of recurrence and prediction of survival time were also investigated but not included in this paper. This paper will only focus on whether or not a recurrence occurs. Although the output remained as a two class model, (recurrence YES or NONE,) two definitions of recurrence were used. Recurrence was defined as at any site or more than local. Applying the ideas initially put forth by Dr. L. Breiman, classification trees and random forests were evaluated as a tool for these predictions. Standard desktop computers using the R2.1 software packages (2005) of lattice, tree and random forest were used for all programming and calculations.

TREES AND PRUNING

Growing A Tree

In this section a summary of the works published by Dr. Breiman (1984) in his book Classification And Regression Trees and his published papers (1999, 2001) on random forests.

The concept of random forest creation requires a basic discussion of tree formation. A tree starts with a seed of collected data with each data point having many variables. One classification variable is selected as the response variable and using the tree, we try to predict the class membership for each data point. The process tries to separate the data into branches and terminal leaves consisting of solitary class membership. These pure class subdivisions are formed using information from the data points. The response variable selected may have a binary or multiple class structure. Each point on the tree is called a node and contains a sub collection of the data points. After answering a question about a variable in each data point, the node is divided (split) into at least two descendent nodes. The new nodes called descendent or daughter nodes contain a partition of the data points from the original or ancestor node. Just as in a real tree a leaf is a direct descendent of some branches but not others even though at the beginning there was a single starting seed. Here too, distal nodes may not be descendents of all the same branches but do have a common starting ancestral starting node.

An attempt is made to split each subsequent daughter node into additional nodes by answering another question. At each split the descendent nodes are to have a higher percentage of data points in the same class. Starting at the original ancestral node and then at each subsequent node the best question about the variables is chosen from the many possibilities so that the answers produce the purest descendent nodes. As the tree grows many end or terminal nodes are present. At each terminal node a different best question will give its best split increasing the purity. Taking the best change in purity from each potential node split, the node split that would produce the best overall increase in purity is the next selected node to be split. Only one node split is performed at a time and then the search for next split begins anew. Increased purity in a node is defined as obtaining a greater percentage of members from one or more classes and smaller number of members of the other classes. Unless other restrictions are implemented, node splitting is not stopped until all the terminal nodes are pure with single class membership.

If we take a sample of data and form this tree so that all of the original data points are correctly classified in pure nodes, a mammoth sized tree can be formed. Especially when starting with a large data set with many variables, it is important to control the behemoth size of the tree. This must be accomplished without significantly sacrificing low misclassification rates. The number of terminal nodes is a measurement of the overall tree size.

Also, we need to test the tree to see if a sample of data defined as the “test set” would be classified correctly as it traversed down (up) the tree? The "training set” is defined as the set of data points used to construct the tree. If the test set is simply a sample from the training set used to form the tree and the tree is left to remain its full size, the accuracy of prediction should be 100%. This value is not predictive for future data points. To get meaningful results a test set must be chosen or ‘set aside’ and kept separate from membership in the training set. This separate ‘out of bag’ set is used for checking the ability of the tree to predict correct class membership. Since the number of data points and input variables are often large a computer algorithm is needed to test the many possible splits at each node.

The following description of the tree formation process will use many of the same notations outlined by Breiman (1984) in his book Classification And Regression Trees.

For the chosen response variable the set of classes may be binary or multiple forming a set C(J) containing the different classifications:

$$\{C(J) \mid J = (1,2,3\ldots j)\}\text{, where J is the number of classes.} \tag{4}$$

There are a total of *N* data points, with each data point *x* belongs to set of data points X:

$$\{ x \in X \mid x = 1, 2, 3....N \}. \tag{5}$$

Each data point entry *x* is a vector made up of *F* input variables:

$$\{ x_f \mid f = 1,2....F \}. \tag{6}$$

Then F will be the size of each data point if all x have F variables in their standard structure even if some data points are missing a value for a variable.

The path of a data point is denoted:

$$\{\mathrm{d}(x) \mid x \subset X\}. \tag{7}$$

Define the ordered pair:

$$(\mathrm{d}(x), j) \text{ where } \mathrm{d}(x)=j, \tag{8}$$

to mean that data point $x \subset X$ passes through the tree along path $d(x)$ and is classified as class j. Let the function representing the ordered pair of the tree classification down a path d(x) and the actual class j, of the data point be:

$$\mathrm{D}(\mathrm{d}(x)= i, j). \tag{9}$$

Then d(x) is correctly classified as i if:

$$\mathrm{D}(\mathrm{d}(x)= i, j)\ [i=j], \tag{10}$$

or misclassified as i if:

$$\mathrm{D}(\mathrm{d}(x)= i, j)\ \ [i \square j]. \tag{11}$$

Let A_j be a subset containing all x where d(x) = j. Then {A} can be a partitioned subset of X into subsets:

$$A_j \text{ where } \{x|\mathrm{d}(x)=j\} \text{ where } j \in A_j \text{ and } X = \cup A_{J_{J=1\ldots j}}. \tag{12}$$

Let the probability that D(d(x) = i,j) be P(i,j). (13)

The success or failure of a classification is the probability of misclassification:

$$(\mathrm{d}(x) = i,j)\ \ [i \square j] \text{ or } \mathrm{P}(i,j) / (\mathrm{P}(i,j) + \mathrm{P}(i,i)). \tag{14}$$

The cost of a misclassification denoted as a, is a non negative continuum of numbers:

$$a \geq 0 \quad \text{occurs when} \quad \mathrm{D}(\mathrm{d}(x)= i, j)\ \ [i \square j]. \tag{15}$$

Let the cost factor *a* also be written as:

$$a = \text{c}(i, j)\ \ [i\square j\ ]\geq\ 0. \tag{16}$$

Define R(d), the indicator misclassification error rate for all (d(x) = *i,* j), [*i* □ *j*]:

$$\text{R(d)} \ = \frac{1}{N}\sum_{n=1}^{N} D(\text{d}(x_n) = i,\, j),\ \text{where}\ D(\text{d}(x_n) = i,\, j) = \begin{cases} 1\ \text{if}\ i \neq j \\ 0\ \text{if}\ i = j \end{cases}. \tag{17}$$

If we use a specific set of $x \in X$ called a 'test set' the equation becomes:

$$\text{R}_{\text{t.s.}}(\text{d}) = \frac{1}{N}\sum_{n=1}^{N} D(\text{d}(x_n) = i,\, j),\ \text{where}\ D(\text{d}(x_n) = i,\, j) = \begin{cases} 1\ \ \text{if}\ i \neq j \\ 0\ \ \text{if}\ i = j \end{cases}. \tag{18}$$

The purity value of the node is a non negative number that inversely decreases as purity of the node increases. It can be demonstrated in the following method. If there are 5 possible classes for all $x \in X$, the purity *i(t)* of a node is how closely all $x \subset t$, have a single classification *j.* If the number of *x* belonging to each class in a node is equally split it is least pure while if all *x* belong to a single class it is the most pure.

Let consider the above example of five classes and define the purity values as:

$$\phi(t) = 1/5,\ 1/5,\ 1/5,\ 1/5,\ 1/5) = i(t)\ =\ \text{some maximum positive value,} \tag{19}$$

$$\phi(t) = (1,\ 0,\ 0,\ 0,\ 0) = i(t) = 0\quad \text{pure node with members of only one class.} \tag{20}$$

The change of purity $\Delta i(s,t)$ going, from a node *t* into two descendent daughter nodes t_L and t_R taking into consideration the probability of each of the two nodes at split s is:

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \tag{21}$$

The combined new purity values *i*(t), of the daughter nodes will be less than the pre-split value of the ancestor node. To grow a tree in the most parsimonious way, look for the greatest increase in purity at each node, or find the greatest $\Delta i(s,t)$ for each split *s* and at each node *t*. At each step in the processes of growing the tree, one searches not only for the best split at each node, but out of all splits in all nodes which split is the best. After each best split there will be a new search at each terminal node for the best possible split. This split would again increase the purity of the terminal nodes the most and therefore lower the purity value *i*(t) the most, i.e. $\Delta i(s,t)$ is greatest.

This process ideally would continue until either all nodes were pure, an arbitrary minimal $\Delta i(s,t)$ was reached, a minimal number of data points were present in each node or a maximum number of terminal nodes were reached. Depending on the data size an enormous tree could be created especially if each node were pure. A 100% correct classification would be obtained if a test set were drawn from the actual training set and then placed into the tree for classification. But what would happen if a new set of data points that was not used in the tree forming the training set was tested in this tree? What would be the misclassification rate R(d) be? Once this mammoth tree was created and placing an 'out of bag' test set revealed a low misclassification rate, our next goal would be to see if a smaller tree would suffice. Limiting the size of the tree can be accomplished by increasing the minimal $\Delta i(s,t)$ needed to make another split, increasing the minimal size of a terminal node or decreasing the maximal the number of terminal nodes.. By utilizing the greatest $\Delta i(s,t)$ available, the most proficient splits would take place.

Another method would 'prune' back the distal twigs after growing a large or maximal tree thus leaving a smaller classifying tree. Using either method, success would be measured by a small tree maintaining or even increasing the accuracy of an 'out of bag' test set class prediction. In this study several patients with similar findings had different outcome classes depending on the size of the tree. Each node split allows more data point facts to come into play determining the correct final class determination. However if we make the tree too large a patient may be placed into the wrong class because of an unimportant single variable used in a split of a small sized branch with only a few remaining data points. This small terminal branch cut can be prevented with limits on tree growth or the making use of the pruning process.

In a table on page 60 of Breiman(1984), demonstrates that as the tree size grows root size to maximal growth, the $R(d)$ for the training set decreases to zero, but if we use an 'out of bag' test set the $R_{t.s.}(d)$ reaches it's minimal value somewhere in-between.

Tree Pruning Process

Instead of stopping the construction of the tree when certain criteria are met, it has been more prudent to allow the tree to grow to maximal size. Then prune back from the terminal node level so that each new tree is smaller then the previous tree in order to reach the proper size tree with an optimal misclassification rate $R_{t.s.}(d)$.

To describe the pruning process a cost factor is introduced into the misclassification picture. Define the cost factor *a* or *c(i,j)* the misclassification as:

$$\begin{aligned} a = c(i,j) \ &\geq 0 \ \ if \ \ i \neq j \\ & \qquad\qquad\qquad \text{where } a \text{ is an arbitrary value } \geq 0 \qquad (22) \\ a = c(i,j) \ &= 0 \ \ if \ \ i = j \end{aligned}$$

The representation of the sum of all the misclassification rates of the terminal nodes $\hat{T}$ is:

$$R(\hat{T}) = \sum_{t \in \hat{T}} r(t)p(t) = \sum_{t \in \hat{T}} R(t). \qquad (23)$$

During the pruning process for each sub-tree T_{n+1} of tree T_n, the size of the $T_{n+1} < T_n$.

Since in the tree growing process the purity of a set of descendent nodes can only increase, the misclassification rate of the ancestral node can only be larger or the same:

$$R(T_{n+1}) \geq R(T_n)\,. \tag{24}$$

The cost complexity of the terminal nodes is defined as:

$$R_a(\hat{T}) = R(\hat{T}) + a\,|\,\hat{T}\,|\,. \tag{25}$$

Where:

$R(\hat{T})$ is the misclassification rate and $\mathrm{a}\left|\hat{T}\right|$ is cost factor times the number of terminal nodes. Looking at a terminal node of the pruned tree T_n, replacing two or more terminal nodes T_t with a single node, the cost complexity of the single terminal node is:

$$R_a(T_n) = R(T_n) + a(1)\,. \tag{26}$$

For the original terminal node branches T_t, its cost complexity was:

$$R_a(T_t) = R(T_t) + a\,|\,\hat{T}_t\,|\,. \tag{27}$$

Since a node split always decreased impurity, for any terminal node the misclassification rate of the pruned branch $R(T_n) > R(T_t)$. If *a* is zero, $R_a(T_n) > R_a(T_t)$ meaning the branch node T_t has smaller cost-complexity than node T_n, but there is some critical value of $a \geq 0$ at which the two cost-complexities become equal. The misclassification rates are discrete values and *a* is a continuum of values. We can find the first node and the lowest value of *a*, that make the two cost complexities $R_a(T_t)$ and $R_a(T)$ equal. From (26 & 27) solving the inequality for *a*:

$$R_a(T_t) < R_a(T)\,, \tag{28}$$

$$R(T_t) + a\left|\hat{T}\right| < R(t) + a\,, \tag{29}$$

$$a < \frac{R(t) - R(T_t)}{\left|\hat{T}\right| - 1}. \qquad (30)$$

As defined above $a > 0$, and therefore the right side of the inequality is positive.

If we have a tree with many nodes $t \in T$, and $t \in T_t$ being the terminal nodes, define a function $g_1(t)$, $t \in T$ by;

$$g_1(t) = \begin{cases} \frac{R(t) - R(T_t)}{|\hat{T}_t| - 1}, t \notin \hat{T}_1 \\ \\ \infty \qquad , t \in \hat{T}_1 \end{cases}. \qquad 31)$$

Setting $g_1(t) = \infty$ when *t* belongs to the set of terminal nodes T_t, ensures that when *a= g(t),* it will prune back a branch that does not represent a terminal node. Define this weakest link and next branch to be pruned as the t in T where:

$$g_1 = \min g_1(t). \qquad (32)$$

This value for $g_1(t)$ at *t* is the weakest link, since as the values of *a* increase, it is the first node at which $R_a(t) = R_a(T_t)$ and becomes the node-split that is removed.

After that node is pruned, the next *g(t)* and subsequent value of *a* is found in order to prune the next node. If two or more branches give the same min $g_1(t)$ then both are pruned. It is also demonstrated on a chart by Breiman (1999) that initial pruning steps actually prune several branches at one time, then as the tree gets smaller, the number of branches pruned at one time decrease towards a single branch. Compared to the tree formation, the pruning process is computationally less time consuming. Depending on the way the test sets and training sets are formed computations for the misclassification values can be listed for different sized pruned trees.

Testing the tree

There are several ways to take a set of data and find a training set to form the tree and then find a test set to obtain the misclassification rate. A few methods will be described using common terminology found words in tree formation literature.

Re-Substitution

Re-substitution takes all the data points as a training set and forms a tree. It then chooses a portion or all of the training set to use as the test set data. Since a tree can grow to total "purity" if we do not prune back the tree at all, a 0% misclassification rate can be obtained. If tree growth is limited or if the largest tree is pruned back the misclassification rates will be higher for the smaller trees. This would be a satisfactory method if prediction of future data points were not a concern. It could be said that there is no demonstrated predictive nature to this tree until an out of bag data set is evaluated.

Set Aside

This method refers to setting aside a random portion of the data to be 'test set' and using the remaining portion as the "training set." The separation of the data into two sets, using only the training set to form the tree and the test set for prediction, corrects the weakness of the re-substitution method. The set aside method is more adaptable if there is an abundance of data available, particularly in comparison to the number of input variables present in each data point. The term 'bagging' refers to the set of data that is used to form the training set. This training set is then used to grow the tree. The remaining or the 'out of bag' data points would then be used as the test set for which the misclassification rates are calculated. Different methods are available for obtaining the random training and test sets. A 2/3 sample set is commonly chosen as the training set to create the tree. That leaves the remaining 1/3 as the test set to determine the accuracy of the classification tree. If the overall data sample size is small especially compared to the number of input variables the test set group may prevent the formation of an adequately sized or diagnostic tree.

Cross Validation

One attempt to correct the problem of limited data size is the use of multiple smaller test groups. Each test set is evaluated using a tree created from a different training set. A random method partitions all the data points into n equal subsets denoted as $V_1,\ V_2,\dots\ V_n$. In turn each subset V_i is set aside as the test set and the remaining subsets are used as the training set to form a tree. Then *n* trees are formed each using the bagged data from the remaining *n-1* subsets. Successively leaving out each of the subsets makes sure that all data points become a member of one test set while being in every other training set. Then each out of bag test set is used in its associated bagged tree to predict its member's classification. Once many trees are created an individual tree may be selected and used for study or all the trees may be collectively used in what is known as a forest.

The misclassification error of a cross validation study is given by:

$$R^{c.v.}(d) = \frac{1}{N} \sum_{\substack{\mathrm{x} \subset \quad V_i \\ \text{where } i=1..n}} (d(\mathrm{x}) = i,\, j),\ [i \neq j]. \qquad (33)$$

The distribution of the test and training sets can be a problem especially when the number of members in each class differs considerably. When forming trees in this study based on classification by recurrence, there were many fewer members in the recurrence group than the non-recurrence group but checks on the percentage in the total test groups randomly selected mimicked the true percentages in the total sample. If the random patient test set selection were not reflective of the total population, computer instructions could be easily given to correct that problem.

Although this paper does not treat regression models, the same methods are available when forming linear equations on training data in order to use on test data.

RANDOM FORESTS

In the above process of the creation of a classification tree, definite criteria are used to grow the tree and perform the node splits. Test sets and training sets are built and used for tree construction and testing according to a defined method. Now with the introduction of random forests, random selection will replace these rules in several steps in the tree formation. Many forms and levels of random selection can be introduced to grow each tree. Randomness at the node level can be of different forms. Examples include:

i. the input variable to select

ii. the data points in the training and test sets

iii. linear combinations of classifiers to try at each node

Breiman (1999) has introduced and written the classic article describing the details of random forests. The basic concepts, some of the mathematical principles and comparisons to other type data models are discussed in this article. His work has shown that for many data sets that were previously studied by other classification methods, the method of random forests has produced the lowest misclassification error rates. The generalization error converges to a limit as the size of the forest increases. Each of the many random trees created will use its d(x,i) to cast a class 'vote' for every data point. Because of the random selection of the tree classifiers, the set of paths d(x,i), will differ for each tree. There is no final single tree and therefore from (9) many D(d(x)=i,j), that is the paths of d(x) in each tree giving a vote for class i are different.

The lack of correlation between these multiple sets of paths help form the predictive strength of the forest. A tally of the class votes for each data point from each tree, determines the final predicted classification of a data point. Majority vote or an alternatively selected percentage can be used to determine the final classification.

In a random forest the key element is randomness. For each tree an independent identically distributed random vector Θ, containing a collection of tree-structured classifiers is chosen. If there are to be K trees then the vectors are denoted $\Theta_1, \Theta_2 .... \Theta_k$. The classifiers are members of the set of input variables. A defined number of input variables are chosen for each node. The variables chosen become the only variables used to find the best split at a node. Each node will use an equal number of independently chosen input variables to find its best split.

Unlike the single tree selection of a set aside sample to use as the test set, the random forest chooses a random training set. The training set and their actual response classes are drawn from the vector set (X,Y), in either a bootstrap manner or without replacement. The remaining data becomes the test set. Irrespective of the method of choosing the training set, approximately one-third of the data will remain as the out of bag test set. Each tree has a different independently chosen random training set. Once the trees are completed and trained, all the data points x in set X, are run down each tree in the forest. Each tree casts a vote for the class prediction of all the data points, but only votes cast by a tree for data points belonging to its out of bag test set are counted. The forest gives a final prediction for each data point *x* after a tally of the votes. The data point's prediction is for the response class receiving the most votes.

Breiman (1999) uses the following notation to define a random forest. For each tree k, the set of classifiers is $h_k(x)$, where data point *x* is an element of X, the set of data points and its class *j* which is an element of *Y* the set of the response classes.
Definition: A random forest is a classifier consisting of a collection of tree-structured classifiers {h(*x*, Θ_k),*k*=1…} where the {Θ_k} are independent identically distributed random vectors and each tree casts a unit vote for popular class at input *x*. The total number of trees in a forest is *k.*

Given the set of classifiers, $h_1(x),\ h_2(x),...h_k(x)$, define the margin function of the set of classifiers as:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j)\,, \tag{34}$$

where *I* is an indicator function. The margin function measures the extent that the average number of correct votes for a data point exceeds the maximum average of incorrect votes for any other class. Strength in the prediction will be reflected by a high margin value. If the margin is negative, the vote is incorrect, i.e. the data point is misclassified. The generalization error is given by the probability over the X,Y space, that the margin function is negative:

$$PE^* = P_{X,Y}(mg(X,Y) < 0)\,. \tag{35}$$

An important basic concept of the random forest is the convergence of the generalization error. The questions of over fitting and reaching a minimum misclassification error that can not be overcome by growing an excessive number of trees, is addressed by

Theorem Brieman (1999): As the number of trees increases, for almost surely all sequences Θ_1, $\Theta_2 \ldots \Theta_n$ the generalization error PE* converges to:

$$P_{X,Y}(P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) < 0)\,. \qquad (36)$$

This convergence means that once a lower level of error over the X,Y space is reached for an expanding forest, continuing to increase the size of the forest without bound will not improve the error rate.

Another form of the margin function of the forest, which is similar to (34) is

$$mr(X,Y) = P_\Theta(h(X,\Theta) = Y) - \max_{j \neq y} P_\Theta(h(X,\Theta) = j)\,. \qquad (37)$$

Here the margin function measures the extent of the probability of correctly classifying *X*, exceeds maximum probability of incorrectly classifying *X*. The probability P_Θ, is over the total set of random tree classifier vectors Θ_n.

The strength and correlation are two parameters that are used for upper bound measurement of the generalization error. The strength refers to the accuracy of the individual classifiers while the correlation describes the dependence between the different classifiers. Their measurements and the relationship are presented next.

The strength of the classifiers is the expected value of this margin function over the X,Y space or:

$$s = E_{X,Y}\, mr(X,Y)\,. \qquad (38)$$

Intuitively, the larger the average margin, the stronger the classifiers. The strength of the classifiers directly relates to the predictive ability of the forest.

For the predictive power of the random forest to be maximal there must be a lack of correlation between the classifying vectors. When the margin of votes is consistently strong and the tree classifiers for each tree are very different and acting independently, a low misclassification rate can be expected. To define a usable correlation measurement, the variance of the margin function is defined as:

$$\text{var}(mr\text{r})= \frac{1}{\text{N}} \sum_{1}^{\text{N}}[(h(x,\Theta)=Y)-\max_{j\neq y}(h(x,\Theta)=j)]^2 \ -s^2 . \tag{39}$$

Breiman (1999) defines the correlation as:

$$\text{corr= var}(mr)/\text{E}_{\Theta}(\sigma(\Theta)^2) \text{ where } \sigma \text{ is the standard deviation.} \tag{40}$$

He further defines the expected standard deviation as

$$(\text{P}_1+P_2+(P_1-P_2)^2)^{\frac{1}{2}} \tag{41}$$

Where P_1 is defined as

$$P_1=\frac{1}{\text{N}}\sum_{1}^{N}(\text{all correct votes}), \tag{42}$$

and P_2 as

$$P_2=\frac{1}{N}\sum_{1}^{N}(\text{all incorrect votes}). \tag{43}$$

A ratio of the correlation to the square of the strength referred to as the ‘c/s2’ ratio. Breiman refers to this ratio as one measure of success of the overall random forest process. A value for this ratio combined with a decreased misclassification error rate indicates a more successful set of classifiers. The increased strength of the classifiers directly relates to a lower misclassification rate while a lower correlation will demonstrate less interdependence of a particular part of a tree. An elevation of *s* and a decrease in the correlation will decrease the c/s2 ratio.

PATIENT DATA

After approval of the internal review board at Duke University Medical Center the data was furnished without any direct patient identifiers. For each of the 1610 patient data points, a list of the input variables was available. A small list of the patients and variables is included in the appendix with the proper headings. Some variables were measured concomitantly with the initial surgical treatment and others during follow-up or after a recurrence. The latter did not figure into the initial treatment risk evaluation. The Duke study collected interval follow-ups on all patients. Follow-up evaluation information is sent to Duke from referring physicians even if a patient is no longer being followed at the Duke Melanoma Clinic. For surgeries performed at an outside hospital, all pathology slides of the lesions are reviewed by the Duke Pathology department. Each patient had a patient number and 22 of their input variables sent to this study. Unknown values of the variables were classified as unknown by protocol.

A partial list of the input variables, codes and explanations follows below. A complete list of the melanoma code protocol is given in the appendix. The PTNum is a patient number formed as a sequential numbering of the patients for the purpose of this study. The AGE is listed in years to one decimal place. The SITEGRP is a non-ordered factor referring to the location type of the primary lesion. The values are 1=trunk, 2=extremity and, 4=head and neck area. A SATEL value of 0=none, 1=yes while, 9=no information refers to the presence of small adjacent satellite skin lesions directly related to the primary lesion. The histological group type (HISTGRP) of the melanoma were condensed and coded as 1=Lentigo Maligna 2=Superficial Spreadinf 3=Nodular, 4=Acral, 5=Other and 6=Unclassified.

The Clark (CLARK) level is a microscopic anatomical tumor depth. It is coded by increasing severity from 1 to 5 equal to the accepted pathologic classification with an additional code of 6 being an unknown Clark level. The Breslow Thickness (THICK) is a micrometer measure of the thickest depth of the lesion. It is given in hundredths of a millimeter and is a continuous variable. This study of thin melanoma limits the patients included to a Breslow thickness of less then 1.00 mm.

The STGGRP represented the stage of the melanoma at initial visit. Its severity varied from single local lesion to distant spread coded as:

0=Primary lesion only (no satellite lesions, no local skin metastsis),
1=’Loc/SatLes’ local skin metastasis or satellite lesions initially,
2=’Intransit’ the presence of intransit nodes (small nodes before the regional nodes)
4=’Nodal’ the presence of regional lymph nodes for example axilla,
8=’Distant’ any distant metastasis.

Ninety-four patients with thin melanomas were noted to have an initial advanced stage.

DAYTOREC is the number of days from initial diagnosis to the first recurrence. Information is coded only for the first recurrence even if is only local and is followed by a more significant recurrence. If no recurrence has occurred the DAYTOREC equals the number of days that has past from initial diagnosis to the last day of follow-up.

RECSTAT is the recurrence status with 1 = no recurrence and 0 = recurrence without consideration of site or severity. The TypeRec refers to the site type of the first recurrence only. TypeRec =0 if there is no recurrence and should be > 0 if RECSTAT =0. The values of TypeRec = 1 for local metastasis, 2 for intransit nodes, 4 for regional nodes and 8 for the presence of distant metastasis. SiteDrec is a measure of the most severe of the recurrences. The different organs are coded differently but their values are not important in this paper. Important values are 30 for non-distant site, 27 for multiple distant sites and 31 for no recurrence. If TypeRec=8 (distant recurrence), a site of distant metastasis should show in SiteDrec scale. Using RECSTAT, TYPEREC and SITEDREC according to the codes gives insight into the difficulty of trying to classify whether or not a patient should be classified as YES or NONE for recurrence. The significance of these difficulties will be discussed later in the paper; suffice it to say for now that a patient with a SiteDrec code of 30 must be looked at carefully when trying to determine if that patient had a predictable and significant recurrence..

LIVE is whether or not a patient was alive at last known contact. It is coded LIVE= 1 (dead) or LIVE = 2 (live). It is important to realize that the vast majority of deaths in these patients are not directly related to the melanoma. DaySurv is the number of days from date of clinical diagnosis to date of last follow up irrespective of LIVE status.

The type case (TYPCAS) equals 1 or 2. The number refers to patients being separated into analytic (1181 patients) and non-analytic (429 patients) respectively depending on whether or not they were seen at Duke during their first round of treatment. If the value is 1 (Analytic), it means that DUMC diagnosed or was involved at the time of the first course of treatment for the primary melanoma. If the value of TYPCAS is 2 (Non-analytic), then DUMC was not involved at the time of treatment for the primary lesion. Typically the Non-analytic patients are referred to DUMC for a recurrence, although they may have been referred just for persistence, progression, or "peace of mind". With this bias in referral patterns, the Non-analytic patients taken as a group are not representative of the outcomes of the general thin melanoma population.

The DEXTENT applies only to TYPCAS = 2 and it is a measure of the extent of the melanoma disease when first seen at Duke. If it is an analytic (TYPCAS=1) patient the value is 0. The DEXTENT values of 1 and 2 are limited to local disease. Codes values of 3 through 6 represent distal disease. A code of 11 means unknown extent or not recorded.

The code ANYIMM refers to the many patients who were also included in an immunotherapy study. A value of 1 means the patient had immunotherapy at some time during the course of treatment, and 0 means the patient has never received immunotherapy.

TREE AND FOREST DATA

The Tree Package (Ripley 2000), allows several options for creating a tree to meet the particular needs of the investigator. The following is an example of the syntax used to grow a tree with some of the parameters used in this study:

```
new.mel.ltr<-tree(response classifier ~ all factors to be considered(separated by +
sign), data.frame,control=tree.control(size of number of training
points,mincut,minsize,mindev))
```

After naming the tree, the tree call is used. First the response classifier variable is chosen, followed by all variables that are to be used as factors. The data frame is then listed. The tree control sets a minimal number of data points (mincut) that must be present in a daughter node. The minimal number of data points in a node that can be divided must be at least twice the value of the mincut. The mindev is the lower limit of the impurity permitted in a node. When that impurity value is reached the node can no longer be split.

Tree Output

To demonstrate an example of a small tree, 1610 patients and 15 variables were used were used to grow a maximally sized tree. The Response variable was typrec where recurrence was defined as any return of melanoma after the initial treatment. The maximal tree was subsequently pruned using a k value of 10. The k value is the cost factor measurement used in the pruning process and it decreased this tree size to 13 total nodes and 7 terminal nodes. The printout of the tree below has the following items included. The ‘split facts’ on the tree show the factor and its values which are the ‘yes’ answer to the question. The column n #pts, is the number of data points still remaining in that node. At each split the number in a daughter node is less than the ancestor node. The first number of the ordered pair yprob, refers to the no recurrence percent and the second number refers to the yes percent. The majority vote is the yval. In this tree the root node 1, is split into nodes 2 and 3 according to the value of dextent with values of 0, 1 and 2 going to the 2nd node and the value of 3, 4, 5, 6 and 11 going to the 3rd node. The proportion of no to yes in the 2nd node is 0.85306 to 0.14694, while in node 3 the no to yes proportion is 0.01140 to 0.98860. This represents a striking change from the root node proportion which was 0.66957 to 0.33043. This was the largest change in purity $\Delta i(s,t)$, available at that split. The split of node 17 into 34 and 35 is based on the factor hist. An asterisk at the end of a line marks a terminal node where the purity can not be significantly improved enough based on the cost factor k. If k were decreased below the present level of 10, further splits would have occurred. The smallest terminal node 35 has 11 patients remaining where as the largest terminal node 34, still has 835 patients.

The Tree

```
                                   * denotes terminal node
node#   split facts          n   deviance        yval     yprob
                          #pts
  1)  root                1610   2043.00     NONE ( 0.66957 0.33043 )
  2)  dextent: 0,1,2      1259   1051.00     NONE ( 0.85306 0.14694 )
  4)  stggrp:  0,1        1196    895.40     NONE ( 0.87625 0.12375 )
  8)  typcas: 1           1118    710.00     NONE ( 0.90340 0.09660 )
 16)  anyimm: 0            272     92.12     NONE ( 0.95956 0.04044 ) *
 17)  anyimm: 1            846    602.60     NONE ( 0.88534 0.11466 )
 34)  hist:1,2,3,6,12,14   835    575.10     NONE ( 0.89102 0.10898 ) *
 35)  hist: 4,10,16         11     15.16     YES ( 0.45455 0.54545 ) *
  9)  typcas: 2             78    108.10     YES ( 0.48718 0.51282 ) *
  5)  stggrp: 4,8           63     85.41     YES ( 0.41270 0.58730 ) *
  3)  dextent:3,4,5,6,11   351     43.75     YES ( 0.01140 0.98860 )
  6)  stggrp: 1,4           12     13.50     YES ( 0.25000 0.75000 ) *
  7)  stggrp: 0            339     13.65     YES ( 0.00295 0.99705 ) *
```

The deviance at each node is a sum of the values for each patient given by the formula:

$$D= -2 \sum_{i=1..j} n_i \ln P_i \,. \qquad (44)$$

Looking at node 8 in the tree above, and applying (66) the deviance = 710.00

$$1118(.90340) = 1010 \text{ patients} \qquad (45)$$

$$D= -2[1118(1010/1118 \cdot \ln(1010/1118)+108/1118 \cdot \ln(108/1118)]=710.042$$

Besides looking at the different size trees and random forests created, several other studies are used to evaluate the predictive results. A misclassification tree plot shows the number of patients misclassified at any node. The table will list the number node corresponding to a node found on the printout of a tree or at a node on the tree plot. Unfortunately the number of data points present at each numbered node that are associated with that number of misclassifications is not given. The terminal nodes are not specifically identified. To obtain the misclassification rate identify the terminal nodes. Then add up the number of misclassified points and divide it by the total number of data points. In the example below the misclassification rate at each node was added to the usual computer output as an additional row. The total misclassification rate for the terminal nodes calculated as above is 10.87%.

```
Misclassification list of the 13 nodes T.N=(16, 34, 3, 5, 9, 5, 6, 7)
misclass.tree  number of misclassified at each node look for n in tree above
node#;          1     2     4     8     16    17   34   35     9     5    3    6    7
misclass #    532  185  148  108    11    97    91    5    38   26    4    3    1
         %    33.0 14.7 12.4  9.66 4.04  11.5 10.9 45.5 48.7 41.3 1.14 25.0 0.30
```

The next output is a list of the deviance values. They are given for each node after a program call for the deviance rates. The order of the nodes is the same as are listed in the tree. The terminal nodes can be identified only by referring back to the starred nodes on the tree.

```
Deviance of the 13 nodes
 2043.04529   1050.93219   895.37925   710.04164    92.12298   602.59383
 575.11749   15.15820     108.07967     85.40603    43.75018    13.49604    13.64905
```

There is a tree pruning call which is used after forming the maximal tree. It uses the k values to set deviance. The tree summary below, lists the proximal most nodes of large sections of the tree that were removed or snipped. All nodes and branches snipped can be seen if the entire tree is printed prior to the pruning process. The variables actually used in the tree construction list only the factors used in the remaining node splits. The other factors considered as possible node split questions but not used are not listed. The number of terminal nodes states the count of the remaining terminal nodes after pruning. The mean deviance is the sum of the deviance values of the terminal nodes divided by the number of data points minus the number of terminal nodes. The misclassification error rate adds the misclassification error rates for the terminal nodes and divides the sum by the total number of data points.

The Tree Summary

```
Classification tree: snip.tree(tree = new.mel.ltr, nodes = 6, 35, 5, 7, 16, 34, 9
Variables actually used in tree construction:
[1] "dextent" "stggrp"  "typcas"  "anyimm"  "hist"
Number of terminal nodes:  7
Residual mean deviance:  0.5633 = 903 / 1603
Misclassification error rate: 0.1087 = 175 / 1610
```

It is important to remember that when growing a tree all misclassification and deviance rates in the tree summary, deviance and misclassification studies above are measured with a re-substitution method not with an out of bag test set.

Cross Validation Deviance and Misclassification Output

A cross validation tree study 'cv.tree' runs the K-fold cross-validation experiment discussed earlier. It will show the deviance or number of misclassifications as a function of the minimal cost-complexity parameter 'k'. In the following a cross validation deviance study is performed on a pruned tree with a k = 8 and 18 terminal nodes. The evaluation of the changes in deviance $dev shows jumps at the $k values below and the number of terminal nodes $size on the top.

```
cv8<-cv.tree(prtree8,K=10)  method is deviance
$size   18 17 16 15 14 13 11  7  6  5  4  2  1
$dev
[1]  966.8582  966.8582  966.8582  966.8582  966.8582  966.8582  966.8582  966.8582
966.8582  966.8582  966.8582 1108.2255
[13] 2045.5177
$k
-Inf   8.087272   8.281956   8.384283   8.645023   8.873169   9.251944   9.727535
12.318136  15.324836   16.605086  73.702425 948.362920
```

A similar cross validation is done noting the number of misclassifications as a function of the cost factor. Here it is noted that the misclassification does not appreciably increase until k is above 11.

```
$misclass
 [1] 189 191 191 193 193 193 195 194 195 532

$k
 -Inf   0.0   0.5   3.0   4.0   5.0   6.0   7.0  11.0 343.0
```

A plot of the deviance and misclassification outputs are available (Figs.1and 2) in the appendix.

Random Forest Formation

The initial decision in tree formation relates to how the test samples are formed. If the entire data set is to be used for prediction the random forest is run using the call name of the data set. The final data set used in the forest is changed to list the input variables as categorical, ordered categorical or continuous valued. The results are listed as class output with the actual and predicted group class memberships the class misclassification rates as well as the total misclassification error rate. The margin values can be easily calculated from the class output for each class. The votes are in general terms without specific mention of particular data point vote.

If testing of a particular sample set is desired with knowledge of the individual data point class membership vote, a random forest is first created with the rest of the data. A prediction sub-routine is then used predicting the class of a set aside test set and will provide the specific vote results for each member of the test set. An adequately sized set aside can be chosen for use in the prediction model if the sample is large enough with few input variables. The test sample may also be chosen randomly and set aside as an out of bag sample while the rest are used as a single large bagged training set. More commonly, the data set is not adequate and there are a large number of associated input variables. Here the bagging type of training set can be difficult to use.

In this data set the problem of the need for an out of bag test set associated with the data sample size available was solved by combining the results of multiple small randomly chosen test sets. Earlier a process of cross-validation was explained. This method was a form of cross validation, but a data point used in one test set was not excluded from repeat use in a subsequent test set.

Breiman (1999) states "There are two reasons for using bagging. The first is that the use of bagging seems to improve accuracy when random features are used. The second is that bagging can be used to give ongoing estimates of the generalization error (PE*) of the combined ensemble of trees, as well as estimates for the strength and correlation." The bias that has been reported for cross validation is not felt to exist in the random bagging procedures. In the computer model algorithm for forming a random forest after the training sets have been used to form the set of classifiers labeled T_K for K=1….k, all data points $x \in X$ are run down each tree; the class is voted on for each T_K. However, the vote for a particular T_K is ignored if the x was a member of the training set that formed its classifiers.

Several other useful choices are available in the R 2.01 random forest subroutine shown below.

```
Rf<- randomForest(response ~ factors (separated by + signs), data set name, ntree=nn,
                                        mtry=zz, nodesize=cc, cutoff=c(aa,bb),  keep.forest=T)
```

Rf is a chosen name for the forest. The first call inside the parenthesis is the response factor whose class memberships are to be decided. The next call is the list of factors to be considered in making the tree each separated by a plus sign. The 'ntree' refers to the number of trees grown in any forest. The call mtry is the number of variables used in each split. The mtry value may range from 1 to the number of factors listed. If the mtry value is called the group size F. The number F is not randomly chosen and stays the same for the entire tree and forest formation. In one study of F values, Breiman (1999) reported on using mtry= 1 and a truncated $\text{int}(\text{Log}_2\text{M}) + 1$, where M is the number of input variables. He added that if many weak variables are present the value of F needs to be increased several fold. His results showed improvements in the convergence, misclassification rates, strengths and correlation values going from F=1 to $\text{int}(\text{Log}_2\text{M}) + 1$. Studies of several different values were made in this paper, which often showed better results when mtry was larger than $\text{int}(\text{Log}_2\text{M}) + 1$, because the variables proved to be weak.

The size of the tree is determined by the selected 'nodesize' which refers to the minimal number of data points that can be present in a node before a vote must be cast and therefore no further splitting is performed. A vote may be cast earlier with a larger number of data points in a node when there are no further good splits available. If the 'nodesize' is decreased a larger tree is formed; although as mentioned earlier the largest tree does not always give the best classification results. Pruning is not performed with random forests.

The cutoff is used when the class membership is grossly uneven and there are not enough members of one class to show reasonable penetration in the prediction results. For the response variable of recurrence, there were only two classes, YES and NONE. Only two cutoff values were used but if there are multiple classes a value is given for each class. Another reason to use adjusted percentages of the cutoff is to decrease the number of false negative votes. In this and many medical data predictions it is more important to identify all recurrence patients at a sacrifice of overcalling the non-recurrence as a false positive.

To easily perform repeat random forest programming studies on the same data sets using different call settings, lettered variables with previously defined but alterable values can be substituted as seen in the above sample. The keep.forest=True is necessary if you either want to reuse the same forest with different settings, build onto the existing forest adding additional trees or use a particular forest for prediction studies on a set aside test sample.

Variable Importance Plot

The variable importance plot (varImpPlot) provides a listing of the variables used in the random forest formation. Each variable is assigned a relative importance value in the final class vote. The importance value for each variable is determined from by calculating the change in the misclassification rate results as a 10% permutation or noise factor is introduced for only that variable. The permuted variable has a higher (stronger) value the more the misclassification rate increases. Neither the manual not the help output for the random forest package (2005), define the exact method of determining this number. As the number of trees in a forest is increased, the numeric values read on the *x*-axis scale for each factor increase but appear to remain in similar proportions. The factors without numeric values when using 300 trees will begin to show a value as the number of trees increase several fold.

METHODS

Initially a maximum tree was grown using all the data and input variables affecting recurrence. The tree was sequentially pruned and studied. The plots of different sized trees are included in the back. Text can be added to any tree but on largest trees it cannot be read and only their plots are shown. A localized portion of a large tree can be visualized using the snip.tree subroutine. As the trees are pruned, the proximal nodes remain the same. To help visualize an entire large tree, a print out listing the tree nodes and a full tree plot is available in the appendix.

The tree(s) formed for prediction are made with a “bagged’ training set. An ‘out of bag’ test set is then used to test the tree. After deciding what limitations to place on the data set, a random sampling of 21 patients was chosen to form the test set. The remaining patients become the members of the training set. The training set is used to form a maximum sized tree that is sequentially pruned to various smaller sizes. Predictions of recurrence on the test sets are performed with the maximum tree and several of the pruned trees. This entire process of tree formation, pruning and predicting is repeated with 5 additional random test-sets for a total of 126 test-set patients in each study. Although the comparisons are not presented, it was found that this number of test-set patients demonstrated to be an appropriately sized representative sample. No attempt was made to prevent an individual patient from being tested twice as it was noted that the training set and tree would be different. A list of the actual and predicted recurrence classes of the individual data point or group class can be programmed in order to compare results. It was possible to pick out all the false positives and false negatives and form a clinical picture of the patients thus misclassified. A list and an intuitive picture of the importance of each input variable used in the tree formation can be extracted from combining the tree summaries for each pruned tree. Several runs of randomly selected groups of 126 patents were made using exactly the same definitions of recurrence and any imposed patient restrictions to evaluate the reproducibility of the results. Closeness of the prediction results appeared related to the comparative percentages of actual YES patients randomly selected for the six test-sets. Results for the single tree were then compared with the random forests created with similar patient and diagnosis restrictions. It is possible for the same test-set samples tested on the single tree to be used and tested as an out of bag challenge to the random forest, but except for the final study, the comparative studies were done on randomly selected samples at the time of the tree and random forest creation.

Repeat runs on the random forests were not necessary after identifying an amount of trees to use that demonstrated reproducibility of results. Studies were performed to find the number of trees which would be necessary to not only give reproducible results, but also make sure that each data point was adequately evaluated. It was also necessary to insure that the vote percentages did not change solely by increasing the number of trees. Studies will be presented showing that a size of 300 trees was the best amount to use. Using more trees, even thousands, did not alter the vote results but did significantly increase the computer time. The stability of a lowest bound of misclassification error in the face of continued increases in tree size is supported by mathematics presented earlier in this paper.

Additional questions arise when growing a forest besides how many trees to place in each forest were

i. What size trees to grow in each forest?

ii. How many variables were to be tried at each node division?

iii. What the best percent cutoff for a YES vote?

The features of trying a different number of variables at each split and cutoff values are not available in the regular singletree formation and pruning program. Random forests results were evaluated to answer the above questions.

Multiple values for the size of a node (nodesize) were used and compared. The nodesize refers to the lower bound of data points that can remain in a node before a class vote must be cast. Letting the number of data points remaining in a node be minimal could provide anecdotal results rather then a most likely scenario. In medical patient results, there is always one case that defies the expected and it is called anecdotal, while it is not to be ignored it should not become a sole predictor as a terminal node.

Different numbers of variables (mtry) were compared for each study. It can be set to any number up to the number of factors being considered in the random forest response list. Starting with the mtry equal to the total number of variables in a particular study, it was then decreased to 12, 10, 7, 4, 2 and 1.

The cutoff which determines the percentage of the votes required for a prediction of a class membership to be made. Because there is a much higher percent classification of actual NONE compared to the actual YES, the terminal node votes were found to be necessarily skewed to the NONE votes. By decreasing the requirement of a cast vote to be YES, the desire was to increase the predictive accuracy of the actual YES group, while also examining what might be an acceptable false positive vote for the NONE group. The majority vote while yielding a low overall TOTAL ERROR rate, had such a high YES M.S.E., that it failed to identify the YES patients that required treatment and was therefore not acceptable and the other cutoff values were necessary. Other cutoff values including .40, .30, .10, .05, .03 and .01 were used. Not all are reported.

For the random forest, tables were made showing results for each node size, cutoff value and the number of variables tried at each node split. For the single tree besides the size of a tree, only the majority vote is supplied with the Tree Package (Ripley, B. 2005). Many other cutoff values were programmed and studied; only several were selected and reported. Unlike the single tree, when evaluating the output of the random forest subroutine, it was not possible to tell the misclassification for any particular patient. Once a random forest is created it was possible to save it and to then test a set aside test sample, getting the predicted class for each test sample data point.

Because there was such a predominance of NONE in any group, using The T.N. which sets the lower bound for the number of data points in a node before the vote must be cast was studied at many levels. The lower the value of T.N. the larger the tree can be formed and the fewer the data points in a terminal node. Also, it is possible to select which factors will be considered as factors for any response variable chosen. In this study the response was recurrence and only initial diagnostic and the use of immunotherapy factors were used. If survival length were chosen to be the response factor many additional variables would have been included as factors.

The first study was performed using all 1610 patients and any input variable that might affect recurrence leaving out factors that measure the timing or type of recurrence. Subsequent studies were done using patient group restrictions and or the definition of what represents a recurrence. Because no information was available on patients who first had a local site recurrence prior to developing a distal recurrence, the presence of a local recurrence could not be used as a factor for distal recurrence prediction. Using any recurrence as significant would mean that the problems of immediate or delayed local spread were considered equal to a distal event. When limiting the patients in a study to those with an initial presence of only local disease, two factors DEXTENT and STGGRP were evaluated. They specifically measured the level of disease at initial diagnosis. Since TYPCAS 1 patients were automatically DEXTENT 0, only a STGGRP less than 2 could be used to evaluate their inclusion. If the patient was TYPCAS 2, a DEXTENT level less then or equal to 2 and a STGGRP less than 2 was required for inclusion. An exception of DEXTEXT equal to 11 or unknown, meant only the STGGRP less than 2 could be used for inclusion.

The restrictions removing patients with DEXTENT greater then 2 or STGGRP greater than or equal to 2 in two patient group studies eliminated patients that had thin melanomas by thickness measure but had an advanced stage of the disease at initial diagnosis by another measure. Some patients already had spread of the disease to a distal lymph node or possibly suffered from a distal metastasis when initially presenting to Duke, even as a TYPCAS = 1 with just the thin lesion. If patients transferred to Duke with advanced disease had full enclosure of their initial data from the outside medical facility, bias would also prevent them from representing an equally distributed thin melanoma patient population.

Further considerations using all factors versus only a limited number of the initial diagnostic variables were necessary. The prediction of recurrence to be made from data at the time of initial diagnosis was found to be highly biased by the strength of the factors regarding the extent and the stage of the disease. The entire patient population was therefore additionally studied not considering the obvious bias of those factors. When using the limited patient population excluding any advanced disease, all variables except typcas were used because of the different types of local extent and stage of the limited disease. Typcas was eliminated as a factor because of the severe bias relating to why a non-analytic patient was referred to DUMC.

Another problem was how to determine what would truly represent a recurrence. If a simple local recurrence at the biopsy sight was the only problem a patient ever had should this to be considered a significant event or should only patients with advanced or distal recurrence be included and thus classified as significant recurrence? Was the local recurrence just related to the type of initial surgical biopsy procedure performed rather than a host-tumor factor? Studies were performed using both definitions of recurrence.

A maximum tree was grown for all the groups above. Each tree was sequentially pruned and studied. Sometimes when creating different sized pruned trees an error in the computer printout meant that a maximum depth of tree was surpassed. It is not clear how the computer worked with these trees when this error occurred and how it would affect the prediction results, so it was decided, without patient selection bias that a completely different run was to be made. Attempts were usually made to find and compare the best prediction results from the single tree and random forest creations.

When forming the maximal and pruned single tree, plots and large tree lists could be visualized in order to see which variable was used to form the split at each node. By looking at the misclassification rate, deviance rate, and the number of terminal nodes present when pruning back a maximal tree, it was clear that a fairly small tree with just a few nodes could be used to produce rates close to the maximal size tree. The particular input variables used to reach these rates approaching the maximal tree were examined and are listed in a tree summary. Graphs and computer lists of the tree results are included in the attached section of printouts and will be cross-referenced in the section dealing with the discussion of the results. The adjustments in the tree formation, which patients and variables to include and how to define recurrence were made after the evaluation of the initial study results. It was only after looking at the findings for all patients and all variables with recurrence initially defined as any return of the tumor did the need for the additional studies become apparent.

Breiman (1999) uses the terms strong and weak factors referring to those factors that demonstrate differing levels of importance in improving impurity levels in the tree formation. Using different plots strong and weak factors can be identified by looking at the changes in the deviance rates and misclassification error rates as the tree is grown or subsequently pruned after full growth. If the major effect on the error rate occurs at the proximal tree with little further improvement in decreasing the misclassification error rate only the proximal factors are considered strong while the remaining are weak. Alternately, at each additional split where error rate and deviance continue to show considerable improvement, the factors involved at those splits are also considered strong.

It was found that the extent and the stage group of the disease were the strongest and dominating factors in the prediction trees but they should not be elevated in the typical thin melanoma patient. For most thin melanoma patients, it is difficult to predict recurrence because they only have only local disease at the initial presentation. If a melanoma patient with advanced disease has a Breslow thickness less than 1.00mm, prediction studies can be performed by removing the extent and stage group factors from any tree classification programs. A search for other strong factors could then be made.

Finding the strength of the variables in the tree formation can also be found by examining the tree structure and associated plots. At each branch on the tree list output the name and value of the split variable, the number of data points in the node, the deviance and the purity of each daughter node is listed. It is possible but tedious to look at the change of these measurements at each step. For the random forests the subroutine local variable importance gives a plot or a list of the factors with their relative importance values.

Previous medical dictum was that the Breslow thickness is the dominant factor to be considered in the severity of the broad range of malignant melanoma diagnosis along with histological type. This patient study focused only on the thin melanoma patient where the thickness was less then 1.00 mm. so it became necessary to re-evaluate how strong a variable is thickness as well as the strength of the other factors in predicting disease recurrence for this set of patients.

In summary the reasons for the many additions to the initial study were;

1) The need to predict a patient's likelihood of recurrence if they only had a thin melanoma meeting the thickness criteria before they were known to have a diagnoses advance disease.
2) The need to identify the combination of factors related to the likelihood of suffering advance disease in a true thin melanoma patient?
3) The patients with a thin melanoma and distal disease have different treatment protocols available then patients with a localized thin melanoma.
4) The strength and bias of the factors DEXTENT and STGGRP.

The six study groups were:

1) All patients, All variables, Any recurrence
2) All patients, All variables, Recurrence more than local

3) All patients, Leave out variables STGGRP & DEXTENT, Any recurrence
4) All patients, Leave out variables STGGRP & DEXTENT, Recurrence more than local
5) Limited patients, All variables, Any recurrence
6) Limited patients, All variables, Recurrence more than local

In numbers 5 and 6 all variables were used because the strong bias of stggrp and dextent to the prediction was already eliminated. The only difference in the patients was the whether or not there was presenting local involvement.

For a tree and random forests error rates outputs include YES , NONE and TOTAL M.S.E. The YES and NONE refer to the class error rate and are reported as a decimal rather then a percentage. Only the TOTAL M.S.E. is reported as a percentage. Similar reporting methods were used in the findings and results section of this paper.

One final test was a direct comparison of the single tree verses the random forest. 6 random samples of 21 patients were selected and recurrence predictions made using the small out of bag samples on both a single tree and a 300 tree random forest. The six samples were tested separately and the results accumulated. In trying to present a fair comparison the parameter of cutoff was set at many different levels in order to find the best results. The MTRY was also varied in the random forest looking for the best results.

FINDINGS

All Patients, All Variables, Any Recurrence

Single trees and random forests were created using all patients, all the available input variables and defining recurrence as any return of the disease after the initial treatment. For this group of 1,610 patients there were 532 out of 1,610 patients or 33.04%

who were actually YES recurrence.

In the growing of a single tree the DEXTENT of disease was the most important input variable and first step in the purification process. The value of the DEXTENT categorical split was 0,1,2 as opposed to patients with an initial DEXTENT showing more than local extension. The next level nodes were separated by whether or not the STGGRP of the disease at the initial presentation had progressed to include intransit nodes, followed by the typcas. The small tree had 4 terminal nodes with a residual mean deviance of 0.5898 or 947.3 / 1606, and a misclassification error rate of 0.1093 or 176 / 1610. It could be formed growing a tree by setting the minimal deviance at .03, or by growing the full tree and then pruning it back using a k value of 30.

Growing a tree with a decreased minimal deviance of .005 or by decreasing k to 12 in the pruning process, produced 7 terminal nodes and introduced the factors anyimm and hist but only decreased the residual mean deviance to 0.5633 or 903 / 1603 and the misclassification error rate to 0.1087 or 175 / 1610. The next variables sequentially introduced were prisite, Clark, side and age but they were not introduced until the number of terminal nodes was increased to 14. To produce this size tree the minimal deviance was 0.038 or the k value was 7.8. The residual mean deviance was decreased to 0.528 and the misclassification error rate decreased to 0.1043 or 168 / 1610.

Interestingly the input variable THICK is not introduced until a tree with 19 terminal nodes using a minimal deviance of 0.00325 or a k value of 7.5. THICK offers little improvement of the residual mean deviance to 0.5072, and the misclassification error rate to 0.09938 or 160 / 1610 for the terminal nodes. This showed that with the re-substitution method increasing the size of the tree and allowing introduction of additional variable factors only decreased the misclassification error from 10.893% to 9.94%. To get a zero misclassification rate and associated zero mean deviance value, a tree with 211 terminal nodes and a k of 1.5 was necessary. The same positioning of variables was found if the tree size was limited by the growth process or by forming a maximal tree and then pruning it back.

Using the findings from the cross validation, misclassification and deviance studies reconfirmed that the major improvement in deviance and misclassification measures occur early at the first node splits. Subsequent node splits had none or only small improvements in error and deviance rates noted with increasing levels in the number of terminal nodes or with decreasing k values. As noted above, the cross validations are out of bag studies as opposed to the re-substitution studies in tree formation but their findings on the misclassification and deviance results were similar.

The tables of results for the misclassification rates using all patients, all variables and with the typerec defined as any recurrence are summarized for the three different tree sizes. The full tree was the same tree that had 211 terminal nodes and a re-substitution zero misclassification error. The midsized tree had a k value of 5, and the maximally pruned tree represented by a k of 30. The percent of actual YES in the 126 sample patients closely paralleled the total group YES percentage. The values for the number of total predicted YES at different cutoff levels are listed on each table. Sometimes the number of predicted YES increased as the cutoff value is decreased lowering the vote necessary to classify a patient as YES from a majority vote to a percentage greater than 15%, 10% or 5%.

Using a majority vote to determine class, the total error rate was 14.29% for the full tree, improving to 8.73% if the k value was 5 and 11.90% for the tree pruned with a k of 30. The M.S.E. for YES was 17.78% for the full tree and k=5 but was reduced in half to 8.89% if the tree pruned fully using k=30. The M.S.E. for NONE was 12.35%, 8.64% and 8.33% respectfully. Decreasing the percent vote necessary to classify a patient as YES affected the M.S.E. for YES and NONE as well as total error. For this study when adjusting the cutoff, the greatest combination improvement in the M.S.E. was for the full tree where the YES improved from 17.78 to 15.56% without decreasing the no vote accuracy and the total M.S.E. improved to 13.49%. With the mid sized and smallest tree the M.S.E. for YES improved to 4.44% and 0% but the NONE M.S.E elevated to unacceptable levels of 45.68 and 100%. The best total M.S.E. for the midsized and small tree was with the majority vote and was 11.90% and 8.73%. The best overall result was for a greatly pruned tree, k=30 had a YES M.S.E. of 8.89%, a NONE M.S.E. of 8.33% and a total M.S.E of 8.73%. However the results for the smallest tree only considered the factors of DEXTENT, STGGRP and typcas.

The same group of patients then had prediction studies performed with the random forest program.

Determining the Number of Trees

To determine the best number of trees to form in each forest repetitive studies were done using different fixed numbers of trees. Several tables are included studying 10 to 5000 trees comparing reproducibility and stability of M.S.E. rates. There was consistency of error rate for a particular number of trees as well as no further decrease in the error rates with increased numbers of trees after levels of 200-300 trees were reached. When using just 10 trees the total number of patients was below 1610, indicating that not all the data points could be evaluated. It would be safe to say that the points reported were not always adequately tested. The studies listed in the appendix were performed with four variables at each node. If two variables were used the results with 10 trees were even worse. In particular looking at the case where 10 trees were used the YES M.S.E. varied from .342-.576 and the total M.S.E. varied from 13.88-21.43%. When 200 trees were used the YES M.S.E. only varied from 0.320 -.338 and the total M.S.E. varied from 11.86-12.73%. At a level of 300 trees the ranges of these error rates were even smaller. Additional tables included show that as the number of trees increased above 300; neither the consistency nor the levels of the M.S.E. results demonstrated any improvement. The findings are consistent with the theory presented earlier that continuing to increase the number of trees in the random forest does not cause over fit. The number of trees for each random forest study was then set at 300.

Determining the Number of Variables at Each Split

When examining the number of variables to try at each split (mtry) often using as many as 13 to 15 variables at each split gave the best results. This might seem to contradict the previous mentioned suggestions that the best number of variables to try at each split is $\log_2 M + 1$. With 15 input variables that would be 4-5. In this case using 13-15 variables meant that the two strong variables would be used at almost every node split. Using so many variables at each split also calls into question the randomness of the tree splitting process.

In the present random forest using a majority vote the M.S.E. for the NONE groups were excellent, consistently in the 5.5-6.5%, when using an mtry of greater than 2 and a larger sized tree. The error increased as the number of variables decreased to 2 and

1. The results consistently showed additional small increases in error rates as the size of each tree size was decreased. This was different from the single tree study. The M.S.E. for the YES group was around 24.5% and varied little with decrease of mtry from 15 down to 4. When the number of variables tried at each split decreased to 2, larger error increases were first noticed and became very large at an mtry of 1. Except for mtry values equal to 2 or 1 the total M.S.E. remained in the 12-12.5%.

As the cutoff was decreased from majority vote to .30, .10, .05, .03 and .01, the M.S.E. for NONE and the TOTAL ERROR increased while the M.S.E. for YES decreased. With the cutoff set at .10 or .05 and the mtry set at 13 or 15, the YES M.S.E. decreased to levels between 12% and 16%, the NONE M.S.E. only increased to 20% and 35% respectively and the TOTAL M.S.E. levels were 18 and 28%. At a cutoff of .01 the YES M.S.E. decreased to 5-6% but the NONE M.S.E. was markedly elevated in the 65-70% range. As the mtry decreased the NONE M.S.E. increased at a faster and to a higher rate. There were only minimal changes noted with a mid valued cutoff of .30. While there was a slight improvement in the M.S.E. YES levels at a cutoff of .30, generally there was not a noticeable change until a cutoff of .10 was reached.

The lowest overall values occurred at larger values of mtry and a cutoff of .49, but the YES M.S.E. was high, around 25%. The best combinations of low false negatives and reasonable levels of false positives occurred with mtry values of 13 and 15 and cutoffs of .05 and 0.1 where the YES M.S.E. decreased to the low teen rate and the NONE M.S.E. was in the 30-45% range. As the number of the mtry decreased, the results were less favorable. They did not appreciably vary with changes in the T.N. size. Whenever the mtry was decreased to 1, there was complete breakdown of the random forest as a predictor.

Using the varImpPlot with the random forest, the number of variables tried at each node was set at 4 and 10 variables and T.N. was set at 2. The factors dextent and stggrp with importance values of .6 and .5 were much stronger than any other factor. They were followed by the factors hist and Clark with values in the .3-.4 range. Satel, race, prisite, ulcer and age followed, all with importance values less then .3. Thickness was listed in the 11th position with a minimal importance. These results are in line with the findings from the single tree formation. Additional studies performed looking at a varImpPlot with different numbers of variables tried and size trees showed little changes in the results unless mtry was 1 or 2 or the largest tree was grown. Also, when using 300 trees or more, the order and relative strength of factors listed remained fairly stable.

All Patients All Variables Recurrence More Than Local

Similar studies were performed with the next group which included all patients and all variables but the definition of recurrence was changed to more than local. With this change only 438 out of 1,610 patients or 27.2% were identified as actual YES. The primary factors in the tree node splits were again dextent and stggrp followed by introduction of anyimm, side, prisite, histgrp, AGE, ulcer and finally Clark and thickness. The first sample of patients had 27/126 yes or 21.43%. The YES M.S.E. was 0.333 for the full tree and k=5, but was a little less 0.296 for k=30. The NONE M.S.E was 0.133 for the full tree and increased to .162 as the YES cutoff was lowered. With a majority vote the NONE M.S.E. was very low, only 0.070 for k=5 and k=30. The levels rose to the mid teen level for the mid level cutoffs. They were increased to .515 and 1.00 for the k=5 and k=30 group when the cutoff was decreased to .05. The total M.S.E. was 17.46%, 12.70% and 11.90% for the full, k=5 and k=30 trees using the majority vote. If the YES cutoff was decreased to .05, there was a slight increase in the total M.S.E. to 19.84% for the full tree but larger increases to 44.44% and 40.48% for k=5 and k=30. A second random sample of 126 patients had 40/126 or 31.75% actual YES. This sample showed some improvement in most M.S.E. results. The YES M.S.E. however were still above .300 for the full tree and .175 and .200 for the k=5 and k=30 trees

The random forest study for this group with the mtry set at 15 showed that for the majority vote the best results occurred in a large tree with a T.N. of 2. It had a NONE M.S.E. of .031, a YES M.S.E. of .299 and a TOTAL ERROR of 10.43%. As the CUTOFF decreased to .05 and .03 the NONE M.S.E. increased to 30% and 40% while the YES M.S.E decreased to 20% and 16% respectively. The TOTAL ERROR increased to 25-35% range. The YES M.S.E. slowly increases as the mtry value was decreased to 7 and 4, but the overall results remained close as long as the mtry remained at 2 or above. At an mtry of 1 most of the actual YES votes were lost until the cutoff decreased to .10. Below that cutoff level the YES M.S.E. was improved to .210-.288. Interestingly the NONE M.S.E. only increased from 0.047-0.138 with a cutoff of .10-.03, a large increase was not noted until after the cutoff was lowered to .03. For all mtry values, many TOTAL ERROR RATE values could be kept in the teen range while having an accompanying 25% YES M.S.E. If an acceptable YES M.S.E would have to be lower, for example in the mid-teens, it was associated with a higher NONE M.S.E. near 40% and a TOTAL ERROR RATE percent in the mid 30’s.

The varImpPlot again showed the dextent and stggrp were the most important variables. Here the thickness had no importance and the importance of Clark also decreased. Age had an increase in its importance value, but none of the other factors while shuffling positions, had any significant change in their relative numeric value. The strengths of dextent and stggrp were still much higher compared to any other factor. Unlike the single tree formation when the factor anyimm had an early penetration, in the random forest no importance was demonstrated.

All Patients, Leave Out Variables, Any Recurrence

Because of the dominance of the two factors dextent and stage group in the full and pruned trees as well as in the random forest, these two variables were eliminated as factors. TYPCAS was also eliminated because it showed up as an artificially strong biased factor related to the reason patients were referred to Duke, rather then with host or tumor biology. A new set of trees were examined using all 1,610 patients first using definition of recurrence as any recurrence. The maximum tree now required 457 terminal nodes to have zero misclassification errors and deviance values.

After the initial split on anyimm, the next two daughter cells split on Clark and thickness. The residual mean deviance was 1.09 = 1743 / 1599, and the misclassification error rate was 0.2534 = 408 / 1610. When pruning the tree with k =11.15 or growing the tree and setting the minimal deviance at 0.005, there were 11 terminal nodes. The factors used were anyimm, THICK, prisite, clark, hist, sex, satel, histgrp and AGE. Decreasing k to 4 yielded the 171 terminal nodes. The misclassification error was still fairly high at 0.1366 = 220 / 1610. Further decreasing k to 3, the number of terminal nodes were 305 and misclassification error rate lowered to 0.06273 = 101 / 1610. This is still very high considering the method of re-substitution is being used to calculate the misclassification rate.

The percentage of patients with any recurrence remained at 532/1610 or 33.04%. In the six samples selected the actual recurrence rate was 41/126 or 33.33%. For the majority vote the NONE M.S.E. was less than 10% but the YES M.S.E. was greater than 60%. As the necessary vote percentage for YES dropped to 15%, the YES M.S.E. decreased to 13.33 % but the NONE M.S.E. increased to 79.01%. If the cutoff was dropped further, the M.S.E. results remained stable. Two different runs were performed with very similar results.

The random forest studies on this group showed that the YES M.S.E. remained elevated above 30% until the CUTOFF was lowered to .05 or below. Unfortunately there was an immediate jump in the NONE M.S.E. increasing to the 75-80% level. There were no significant changes (improvement) noted by changing the T.N. or mtry. Setting mtry at 7 or 4, the varImpPlot listed prisite as the strongest factor followed by hist, Clark and race. Anyimm, thick , satel and ulcer were of minimal importance. These again were different from the single tree study where anyimm, thickness, prisite and Clark were the first split variables.

All Patients, Leave Out Variables, Recurrence More Than Local

Using the same patient and variables the typerec was changed to more than local disease. The tree formation showed the dominant factors were anyimm, Clark, hist, AGE and satel. Interestingly, THICK again entered late but was the split factor for one of the daughter cells of the primary node. There were 442 terminal nodes necessary in the full tree in order to get a zero misclassification and deviance rate. With a majority vote the NONE M.S.E. was near 14% but the YES M.S.E. was very high at over 70%. Dropping the YES to .15 and below, there was dramatic shift with the YES M.S.E. being only

5.56% but the NONE M.S.E. rose to 73-76%. Several cutoff step studies were done lowering the vote percentage necessary to give a YES, but there was not a value where the YES M.S.E. dropped without a concomitant rapid increase in the NONE M.S.E.

For the random forest creation there were no combinations of mtry, T.N. and cutoff which gave either low YES M.S.E. or combination low NONE M.S.E. values with reasonable levels of the companion YES M.S.E. to give meaningful results. The varImpPlot showed that the most important factors were prisite, age, hist and race. The other factors had a minimal effect.

Limited Patients, Leave Out 'typerec', Any Recurrence

The last two groups involved using a limited number of patients leaving only the patients with stggrp less than 2 or dextent less than or equal to 2. Initially the typerc was set at greater than zero. There were 1,219 patients remaining with 13.95% or 170 patients classified as YES. The full tree had 189 terminal nodes. As the k value was changed the first two splits were for typcas and dextent followed by prisite and hist. The YES M.S.E. was elevated around 50% and did not lower until a mid-sized tree with a YES cutoff at .05 was examined. However the NONE M.S.E. remained low not rising above 20.18% until the YES cutoff of .05 was reached in the mid-sized and smaller tree when suddenly it increased dramatically. Whenever the YES M.S.E. decreased to a reasonable amount, the NONE M.S.E. level increased dramatically.

Like the single tree the random forest also could not manage to find levels where the YES M.S.E. can be lowered without having a very high NONE M.S.E. The YES M.S.E could be lowered to 25-35% range but with a 50-60% NONE M.S.E. A varImpPlot showed that the dextent was by far the most important followed by hist, Clark, age and race.

Limited Patients, Leave Out ‘typerec’, Recurrence More Than Local

The last of these six studies defined recurrence as more than local for the limited patient group. There were only 114 out of 1,219 or 9.35% who were actually classified as YES. Neither the tree nor the random forest displayed any ability to predict recurrence. The varImpPlot sequentially listed dextent, hist, ulcer and race as most important. Except for dextent, all factors were weak with values less than .25. The Thickness and Clark were minimal factors.

Direct Comparison Random Forest and Single Trees Same Patients

The final study used a sample of 126 patients and made a direct comparison of the predictive ability of the random forest versus a single tree of different sizes. It showed that if all patients and all variables were used and recurrence defined as any recurrence, the random forest could produce a better YES M.S.E by over 10%, but the overall M.S.E. and in particular the NONE M.S.E. were adversely affected by approximately 20% and 40% respectfully. If recurrence was defined as greater than local, the YES M.S.E. improvement of the random forest over the single tree was closer to 5% while the other spread of the other M.S.E results increased slightly.

Leaving out the strong variables of stggrp and dextent and using any recurrence, the NONE M.S.E. was very high at 70-80%. The TOTAL M.S.E. was between 50-60%. The YES M.S.E. was 7.69% for the full and mid sized tree compared to 14.29-15.38% for the small tree and random forest. Neither system was able to produce usable predictions. Many values of mtry and cutoff values for both single tree and the random forest were tried without any improvement in the results. With recurrence defined as more than local, the results were very similar. If the variable typcas were not eliminated the rates were much better. However as mentioned above there is a bias introduced because most of the patients in the non-analytic group had reasons for being referred that were reflected in the stggrp and dextent values. Therefore these results are not included.

Using a limited number of patients, all variables and any recurrence the YES M.S.E. was 31.58% and 22.22% for the two runs with the NONE M.S.E. at 46.73% and 34.26% where as with the single tree the YES M.S.E. was much higher in the 50-83% range but with NONE M.S.E. rates close to single digit levels. Using recurrence defined as greater than local the YES M.S.E. for the random forest was 33.33% and 40% but the NONE M.S.E. were 74.77% and 66.67%. This represented an indiscriminant prediction of patients as YES. The single tree again had low NONE M.S.E. rates but 60-80% YES M.S.E.

Evaluation of Correct and Misclassified Limited Patients

The most important group of patients are the ones with no advanced disease at initial presentation and with recurrence defined as more than local. 1219 are in this group with 114 or 9.35% classified as actual recurrence. A random forest was created with 500 trees and 4 variables at each split. A cutoff of .75 and a node-size of 2 were used. Plots and summaries of some of the variables are listed as items in the appendix.

The TOTAL, NONE and YES M.S.E. are 53.97%, 57.66% and 26.67% respectfully. The summaries of the entire group of patients in the limited group showed similarities in percentages in AGE and sitegrp. There was an 8.62% higher percentage of females (sex=2) from the entire population who qualified for the limited subset than compared to the percentage of males. Of those patients in the limited population, there was a 2.43% lower percentage of females compared to males who subsequently had a recurrence. The histgrp showed an imbalance as 21.62% of the acral, pagetoid and 'other' pathology types in the YES subset. There was not an appreciable difference in the Clark levels. Looking at the qqplot of thickness (Item 1.), there was less than a 0.1mm increase in thickness at the percentile level for most of the YES group. Looking at the use of immunotherapy, 91.23% of the YES group compared to 74.35% of the NONE group received therapy. This is stated without looking at the statistics of which patients were chosen to receive immunotherapy.

Next comparing the actual YES patients who were correctly verses misclassified. The age of the misclassified was 10 years older at each percentile. A higher percentage of the females were misclassified. Half of the extremity sitegrp was misclassified. Only one low level Clark 1 and 2 was misclassified. All Clark levels greater than 3 were classified correctly. The mid Clark had 3 of 9 patients misclassified as no recurrence. The thickness qqplot (Item 2.) showed that the correctly classified patients had thicker lesions. It is very interesting that the random forest was able to correctly classify the two patients that were Clark level 1 as YES. The two patients that did not receive immunotherapy were both misclassified as no recurrence.

For the actual NONE patients 57.67% were misclassified. A profile comparing those who were misclassified verses those correctly classified showed that the thickness was 0.1-0.2 mm. thinner above the 25 percentile level (Item 3). The AGE of the misclassified was 5-10 years below the correctly classified in the 25th-75th percentiles. Females were misclassified as YES at a much higher rate 71.2% than the males 44.4%. The histgrp superficial spreading was most often misclassified as yes. The superficial spreading melanoma generally has a better prognosis and 5 year survival than the nodular type. For some reason this is not being recognized by the forest. Also patients that received immunotherapy were often misclassified as yes.

RESULTS

When looking at the results for most of the studies, there is an apparent failure of the random forests to produce the high degree of correct classifications that the theory and past experiences of other researchers have found. We feel this does not represent a fault of the tree or random forest, but instead is caused by lack of correlation and strength of the data factors allowing recurrence predictions.

In the first group, while reasonable percentages were reached when including all patients, all the variables and any recurrence, the dominance of the initial stage and extent of the disease question the validity, meaning and significance of these findings. The misclassification rate of the single tree using re-substitution was a low10.9% with use of only four terminal nodes using just the factors dextent and stggrp. After finally introducing the eighth factor, thickness, the misclassification was only reduced to 9.9%. The only strong variables found in the random forest variable importance plot were dextent and stggrp. In this study the results for the random forest were better than for the single tree. Using a cutoff of .1 or .05 a 12-15% false negative rate was accompanied by a false positive rate of only 20-30%. This would represent treating only 40% of all patients instead of all patients missing less than 15% of patients that might need additional therapy.

When recurrence was more appropriately defined as more than a local recurrence, the single tree had a YES M.S.E. above 30% for the full tree and around 20% for the smaller trees with the NONE M.S.E. in the teens. This would unfortunately miss treatment for too many patients that will have significant recurrence. With the random forests a YES M.S.E. could be decreased to the 10-20% range but the NONE M.S.E. increased to the range of 40-50%. With these results about two thirds of the total patient population would be treated, missing an average of only 15% of the patients that need treatment.

If the only strong factors dextext and stggrp are removed from either tree process it is not surprising to find that predictive results are poor. Now the first factors introduced were whether or not a patient received immunotherapy, followed by Clark level and Breslow Thickness. Neither the single tree nor the random forest had any ability regardless of tree size, number of variables tried or cutoff points chosen to have an acceptable YES M.S.E. without an unacceptable NONE M.S.E. These results were due to the lack of no new strong factors found in the variable importance plots. Had they been present in this study, some strength should have been exhibited in the prior study.

When the patients with advanced disease were removed from the study the tress and forest predicted recurrence equally in patients with actual recurrence as no actual recurrence. Using majority vote the YES M.S.E. was too high to have any use in trying to decide which patients should get adjuvant therapy. When decreasing the percentage of votes necessary for the prediction of either a single tree or the forest to classify a patient as recurrence (YES), the error rate of the YES vote decreased but the accompanying false positive error rate of the non-recurrence rose rapidly. Therefore there were no levels of the cutoff that could be identified capturing a lowering of the false negative rate but without the sudden rise in false positive rate.

Stability and reproducibility of the random forest predictions were shown at 300 trees with the 15 input variables and 1610 patients. Less than 300 trees showed inconsistency in percentages of error rates and use of more than 300 trees did not produce any further improvement. Results using mtry set at more than 4 usually showed increasing results. Using an mtry of 2 gave poor results, but using a single variable at each split as is routinely done in single tree formation, often gave a complete breakdown in error rates.

CONCLUSIONS

The purpose of this thesis study is to try to predict which patients with a thin melanoma are at risk for recurrence using the methods of trees and random forests. The ideal would be to treat only those patients who will suffer a recurrence and not treat those who although have the same diagnosis will not suffer a recurrence. The intent was to use artificial intelligence to compile a picture from 1610 patients with a thin melanoma and using a multitude of input data variables predict which patients will have a recurrence. Besides providing patient with information there are several major positive results in forming such a mapping:

1) Only expose the highest risk patients to potentially dangerous adjuvant therapy
2) Have a positive affect on cost-benefit analysis of such therapy
3) Have a better understanding on the success of the therapy if the use is limited to the high risk patients.

When all patients were entered into the study and the definition of recurrence was any recurrence, at the best cutoffs there was a low false negative error capturing 85-90% of the patients who ultimately recurred and a reasonable false positive rate of 21-34%. Using the 90% of YES patients and the 34% of NONE only 808 of 1610 or 50% of the patients would undergo treatment. If recurrence was defined as more then local the best results showed a 16-20% false negative and a 30-40% false positive rate. This would mean that 869 of 1610 patients or 54% would be treated. Unfortunately once either the major factors of initial extent and stage disease of the patients disease were eliminated as factors or patients with advanced disease were removed from the study, the results were poor and essentially of no clinical value. Additional studies can be performed to profile the patients misclassified looking for patterns in their data. In e-mail correspondence with Dr. Breiman (2005) about this apparent failure, he wrote “Then you will have to solve the mystery by being your own detective. I will say that (properly set up) Random Forests are almost always representative of the information in the data.”

What then is the answer to this mystery as to why correct predictions cannot be made? Assuming the data is collected accurately is there adequate information available? Certainly the data about a patient will never be complete, but it can frequently be updated. Additional information about each patient is being accumulated which may affect the random forest’s ability to make predictions. The difficulty in the addition of more variable factors into the study will be a minimal programming challenge. The computerized data banks can add a column to each patient record that can then be read into a program written using the R2.01-RandomForest package (Breiman 2005), without difficulty. Although it is not known which factors might be strong variables especially when combined with the other variables, information about the type of original biopsy, any microscopic findings of mitosis per high power field and the presence of microvascular or lymphatic invasion seem the most promising. Neither would be related to the gross extent or stage of the disease nor affect the classification of thin melanoma. There are other heredity, social and other personal data facts being collected that will be inputted into the study. One can only surmise as to their actual importance. Gene mapping or expression is certainly in the future.

The methods of trees and random forests are examples of machine learning and artificial intelligence. Quoting Douglas Hofstadter (1979), a list of several statements worth repeating, reflect on the essential abilities and necessities for intelligence.

i. to respond to situations very flexibly

ii. to take advantage of fortuitous circumstances

iii. to make sense out of ambiguous or contradictory messages

iv. to recognize the importance of different elements of situation

v. to find similarities between situations despite differences which might separate them

vi. to draw distinctions between situations despite similarities which may link them

vii. to synthesize new concepts by taking old concepts and putting them together in new ways

viii. to come up with novel ideas

A few of these statements create a paradox when dealing with computer based artificial intelligence, because by their very nature computers are the essence of unconscious automatic premeditated decision making. This study has been another attempt to use multiple combinations of data making use of a large patient database to make intelligent machine predictions in patient response to a disease.

Even if all the data were available, how would it compare to the excellence of the highly trained treating physician? When a doctor practices the art of medicine, he or she must gather all the information available and mix it together along with a sometimes, intangible feeling of why a patient might behave or must be diagnosed in a manner that contradicts the statistical nature of their data. The correct diagnosis and treatment of an individual patient sometimes defies all the numbers collected in a computerized paper file. A perfect machine would have therapy limited to only those that need and could benefit from treatment but eliminate from receiving treatment, those who cannot benefit. Is there ever room for the necessary flexibility required for the patient that does not fit the mold?

Finally, there is no limit to the cancer types and stages to which one may apply the theory of trees and random forests. Questions such as which patients should be treated and with which treatment, abound in medicine. The use of Random Forests in thin melanoma might not have been the initial best choice of disease because of the results found. However, Dr. Seigler and D.U.M.C. was gracious enough to share their data so that an experience and exploration of the use of random forests in treatment protocols could get a start.