Predicting 5-Year Survival of Colorectal Carcinoma Patients Using Data Mining Methods
David Chhieng, MD¹, Michael Hardin, PhD², Billie Anderson, MS², Upender Manne, PhD¹,
¹Department of Pathology, University of Alabama, Birmingham AL and ²Department of Information Systems, Statistics and Management, University of Alabama, Tuscaloosa AL

Abstract
We compared the accuracy of 3 data mining models, neural network, decision tree, and logistic regression, in predicting the 5 year survival of patients with colorectal cancer (CRC). The database consisted of patient demographics, pathologic features, and levels of expression of 2 biomarkers (p53 and Bcl-2). All 3 methods demonstrated acceptable accuracy, ranging from 64% to 70%. The neural network model had the best specificity (80%) and accuracy (70%) but lowest sensitivity (59%). Both logistic regression and decision models demonstrated comparable sensitivity (72%).

Introduction
Predicting clinical outcome in patients with CRC would allow clinicians to offer customized treatment protocols. Currently, predicting outcome in cancer patients relies on statistical survival analyses. Few studies have examined the use of data-mining techniques for predicting outcome in patients with CRC. The objective of this study was to develop prognostic models using data mining tools to predict the 5-year survival of patients diagnosed with CRC.

Materials and Methods
The database consisted of patient demographics (age, gender, and ethnic background), tumor characteristic (size, stage, and degree of differentiation), and phenotypic expression levels of 2 biomarkers, Bcl-2 and p53, of 491 patients who were diagnosed with CRC from 1981 to 1993. The software package utilized was XLminer (Resampling Stats, Inc. Arlington, VA) an add-in for Excel (Microsoft Inc. Red Mountain WA). Three classifications models: neural network, decision tree, and logistic regression were used to compare the accuracy of these classification models.

Results
Table 1 summarizes the operating characteristics of the 3 models. Based on the logistic regression model, age ($p = 0.05$, odd ratio = 1.943), the tumor stage (T) of primary tumor ($p = 0.005$, odd ratio = 1.942) and the presence and absence of metastatic disease ($p = 0.002$, odd ratio = 25.800) emerged as the 3 most significant variables. Based on the decision tree model, the variables remained after pruning included patient age, tumor stage (T) of the primary tumor, the presence or absence of metastatic disease, and p53 status. The ANN consisted of 1 hidden layer and 25 nodes.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>66%</td>
<td>72%</td>
<td>61%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>64%</td>
<td>72%</td>
<td>56%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>70%</td>
<td>59%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Table 1 Operating characteristics of the 3 models

Conclusions
All 3 methods demonstrated acceptable accuracy with ANN reporting the best performance in terms of accuracy and specificity. Both the decision tree and logistic regression models demonstrated similar sensitivity. Additional studies with larger patient cohort and variables would further improve the performance of these models.

These studies are supported by a grant from NIH (RO1-CA098932).