A Novel Use of Data Mining to Predict Outcome of Oropharyngeal Squamous Cell Carcinoma

James Martin MD, Mary Malliaris PhD, Swati Mehrota MD, Sarah Wirth BS, Maria Strus BS, Carol Bier-Laning MD, FACS
1Department of Otolaryngology, Loyola University Medical Center, Maywood, IL 2Quinlan School of Business, Loyola University, Chicago, IL 3Department of Pathology, Loyola University Medical Center, Maywood, IL 4Stritch School of Medicine, Maywood, IL 5Loyola University, Chicago, IL

Introduction
The incidence of oropharyngeal squamous cell carcinoma (OPSCC) has been increasing over the last two decades relative to other head and neck cancers, thought to be in large part due to the finding that a large percentage of OPSCC is associated with Human Papilloma Virus (HPV). HPV associated OPSCC is seen more often in younger male patients who may not have the traditional risk factors of head and neck squamous cell carcinoma, namely tobacco and ethanol abuse. HPV associated OPSCC also appears to show differences in response to treatment and prognosis.

Traditionally, cancer staging, taking into account tumor size, extent of nodal disease and metastasis was the primary or only method used to determine prognosis and guide treatment planning. A recent review published in Cancer, showed that TNM staging was predictive of OPSCC survival in cases diagnosed between 1995 and 1995, but after 1995 TNM staging had no significant correlation to survival. This suggests that other tumor and patient factors may be more important in understanding prognosis, particularly in HPV related OPSCC.

We were interested in discovering what other factors may be useful in predicting outcome in OPSCC. We chose to use data mining strategies as an initial step to determine factors that may be worthy of further study.

Unlike traditional statistics in which a sample is drawn from the population of interest and tested to determine whether or not it supports the null hypothesis, data mining is a strategy that uses all of the data in the complete data set to look for associations and patterns. It is especially useful and powerful, because unlike traditional statistics it is designed to analyze vast amounts of data with multiple interrelated variables.

Methods
This was a retrospective study of 124 patients with OPSCC treated at Loyola University Medical Center from 2000-2009 for whom 2 year outcome was known. Patient and tumor data, including TNM and overall stage, type of primary and salvage treatment, persistence and survival were collected from the medical chart. HPV status was obtained by immunohistochemical staining of archived tissue blocks for p16. IBM’s SPSS Modeler data mining software used to build a decision tree model and analyze the model. Microsoft Excel 2010 was used to construct the pivot tables. The target outcome was 2 year survival.

Patient characteristics
- N=124
- 75% male
- Average age 58
- 69% smokers, 22% alcohol abuse
- 57%, BOT, 40% tonsil, 2% soft palate, 2% post O/P wall
- 71% HPV positive
- Stage distribution
  - I=4
  - II=7
  - III=21
  - IV=90
- Primary Surgery 80%
  - 75% followed by adjuvant RT with 23% also receiving chemo
- Primary CRT 19%
  - 46% with salvage surgery
- One patient with RT alone
- Two year survival 85%

Results

Figure 1: Decision Tree
The schematic representation of the generated decision tree. This tree shows the rules created to optimally predict outcome from the training data set. The validation set was then tested using these rules.

Figure 2: Decision Tree Predictor Importance
Graphical representation of the relative importance of the input variables in the formation of the decision tree. Note that COAD, age at diagnosis and smoking status are all considered relatively more important by the model than HPV status.

Figure 3: ROC Space Plot
ROC Space Plot comparing the performance of the training (TR) and validation (VAL) sets for 3 different prediction methods of two year overall survival. Decision Tree (DT) used the generated decision tree analysis from Figure 1 to predict outcome. HPV uses only HPV status of the tumor to predict outcome and Stage IV uses only Stage IV status to predict outcome. Points along the red line are equivalent to a random guess; those along the line are better than random; those below the line are worse than random.

Figure 4: Pivot Tables
The upper table shows the relationship between HPV status by row versus 2 Year Overall Survival by column. The lower panel shows the relationship between Surgery as Primary Therapy by first order row and HPV status by second order row versus 2 Year Overall Survival by column. Note the survival with primary surgery in HPV positive patients (89%) compared to nonoperative treatment in HPV positive patients (78%).

Conclusions
Data mining offers a novel approach with a strong potential for predicting outcomes and guiding therapy using multiple patient and tumor characteristics. Our study shows that data mining utilizing multiple patient factors provides better outcome prediction than a stage based model. Further refinement of the model with additional tumor information and complete, prospective data collection is planned to improve outcome prediction.

References