A Bayesian Network Model of Head and Neck Squamous Cell Carcinoma Incorporating Gene Expression Profiles

Shuyang Wu

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Fredric Wolf, Chair

Mark Whipple

Mark Phillips

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

A Bayesian Network Model of Head and Neck Squamous Cell Carcinoma Incorporating Gene Expression Profiles

Shuyang Wu

Chair of the Supervisory Committee:
Professor: Fredric Wolf
Biomedical Informatics and Medical Education

Radiation therapy is a treatment for metastatic Head and Neck Squamous Cell Carcinoma, which allows precision targeting of certain groups of lymph nodes. A Bayesian network predictive model was developed aiming to help achieve such precision using information on the primary site and size of the tumor, representing the current decision making process in clinical settings. Additional risk factors, the patient's genetic profile and smoking history, were added to examine their predictability of metastasis through the improvement in prediction accuracies. The model was trained with publicly available data extracted from the Cancer Genome Atlas (TCGA) and validated against the TCGA dataset as well as clinical data reported to the University of Washington Tumor Board. Results show that genetic profile data improves model accuracy and

such improvement may affect clinical decision making especially for patients with more advanced metastasis. A prototype for decision support application was built based on the results to demonstrate the clinical significance of the model. However, more data is needed to show significance of the proposed effects, as well as to improve accuracy of the overall model.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Mark Whipple for his guidance and continuous support of my thesis project, also to my committee chair, Dr. Fredric Wolf, and my committee member, Dr. Mark Phillips, for their insightful comments and feedbacks. Besides my committee, my sincere gratitude goes to Dr. Ira Kalet, whose work is the foundation of this study. Additionally, I thank Anthony Law, Lucy Wang and Hyunggu Jung for the valuable contributions they offered while the study was being conducted.

CHAPTER 1 INTRODUCTION

Head and neck cancers account for approximately 3% of all cancers in the United States [1]. More than 90% of head and neck cancers are squamous cell carcinoma [2]. Head and neck squamous cell carcinomas (HNSCC) have broadly varying survival rates, depending on the primary site, disease stage and the occurrence of metastasis [3]. HNSCC initially metastasizes to the lymph nodes in the neck, following lymph drainage pathways. The regions of the neck containing lymph nodes are classified into six imaging-based surgical neck levels I through VI as shown in Figure 1[4][5]. If PET scans or CT scans show evidence that the cancer cells have spread to any of the lymph nodes in a level, radiation therapy can be targeted to treat the level of interest. Radiation therapy is also prescribed if the tumor has reached a certain size, even if there are no detectable signs of lymphatic metastasis. This is because with current technology, small amounts of cancer cells in the lymph nodes do not appear on scans, but there is enough risk in these areas to warrant treatment. Being able to predict specific locations of lymphatic metastasis is critical for both minimizing the risk of recurrence and minimizing the complications resulting from unnecessary radiation.

Since not all lymph nodes are equally likely to be involved in metastasis, physicians determine which lymph nodes to target based on prior knowledge and personal experience. The decision making process requires them to estimate many variables such as which lymphatic channels the tumor cells have taken and how far along the channels they have spread [6]. A study by Crosskerry showed that physician judgment can vary from reality by 15% on average [7]. In other words, there is approximately 15% probability that physicians might make either over-conservative decisions and treated the healthy area or leave the cancerous lymph nodes untreated. There have been some studies that examined the likelihood of certain groups of lymph

nodes having cancer given the prior state. For example, a hidden Markov model was developed based on lymphatic anatomical structure, the primary tumor location and T-stages [6]. A predictive model applying the Bayesian network approach was also established with the same predictors and evaluated to show clinical significance, meaning it could successfully improve accuracy for medical decision-making [8]. With the promising results of both predictive models, we suspect that additional information could be added as predictors to improve model performance.

Studies have shown that there is a genetic expression profile predictive of nodal metastasis of oral cancer [9] [10]. The proposed profile is identified through differential analysis of Affymetrix Human Genome Focus arrays and confirmed by immunohistochemical analysis for transglutaminase-3 and keratin 16 [9] [10]. We hypothesize that integrating genetic profile information into the predictive model will improve its performance, leading to more accurate decision aids for clinicians. Also, we will be able to quantify the effects that genetic profile have on metastasis by comparing the performance of models with and without it. On the other hand, although tobacco smoking and alcohol consumption are the leading causes for HNSCC [11], they are unlikely to affect lymphatic metastasis. This assumption can be investigated as well by incorporating the smoking history as an additional predictor into the model.

We propose to extend existing predictive models to capture the probabilities of finding cancerous lymph nodes in each neck level using patient's primary tumor site, tumor size and genetic profile (or smoking history indicator). Such a model could potentially help physicians to make improved evidence-based decisions while performing targeted treatments such as radiation therapy. The studied effects of a genetic profile on metastasis may enable physicians to adopt more individualized treatments.

CHAPTER 2 METHODS

Two Bayesian network models with different structures were constructed, one used for evaluating the effects of adding additional predictors, the other used to make a prototype for the decision support application. I will refer to these two models as the evaluation model and the prototype model in this paper. We used two data sources to train and test the models. Data on the primary site, genetic profile, and metastatic levels were be pre-processed to be better incorporated into the two models. Parameter learning and cross-validation results were generated to evaluate model performance and build the decision support application prototype.

2.1 Model Structure – Evaluation Model and Prototype Model

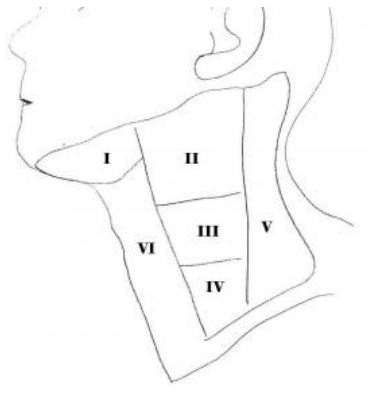


Figure 1. Regions of the neck classified by the surgical levels I through VI. Level I, submental and submandibular group; Level II, upper jugular group; Level III, middle jugular group; Level IV, lower jugular group; Level V, posterior triangle group; Level VI, anterior compartment.

To model the nodal metastasis of HNSCC, we need to understand the anatomical structure of the lymphatic system in the neck. There are six surgical neck dissection levels, shown in figure 1, which are used to delineate regions of cancerous tissue targeted for treatment. Nodal metastasis tends to follow well-delineated pathways that map to the regions described by these surgical levels [12]. Based on the physical connectivity of the lymphatic system, cancer cells can only travel to lymph nodes that drain the primary tumor, and to other lymph nodes through connected drainage channels. We were able to draw detailed anatomic knowledge on

lymphatic connectivity from the Foundational Model of Anatomy (FMA) [13]. We assume that despite primary sites or T-stages, once the cancer cells enter the lymphatic system, the metastasis pathway remains consistent and unidirectional, which means the spread of cancer occurs only along described lymphatic channels, and in a direction away from the primary tumor.

2.1.1 *Evaluation Model Structure*

For the purpose of studying model performance with additional predictors, we made a simplified assumption that the spread of cancer follows a linear fashion from level I to level V. Clinically, there are exceptions due to skip lesions and branching, but representing these scenarios can complicate our interpretation of adding additional predictors. Therefore, we developed a Bayesian network model with the structure shown in Figure 2. The model outcome is generalized to indicate the highest level of metastatic prognosis. This means for each level below the highest, the estimated probability of that level containing cancer cells is the sum of the probabilities of all levels at or above it. The inputs of the baseline model, based on existing research and our hypothesis, are the tumor origin, the size and local involvement of the primary tumor represented by T-stage. The patient's tumor-associated genetic profile and tobacco smoking history indicator were added as inputs into the baseline model separately to study their effects on model performance.

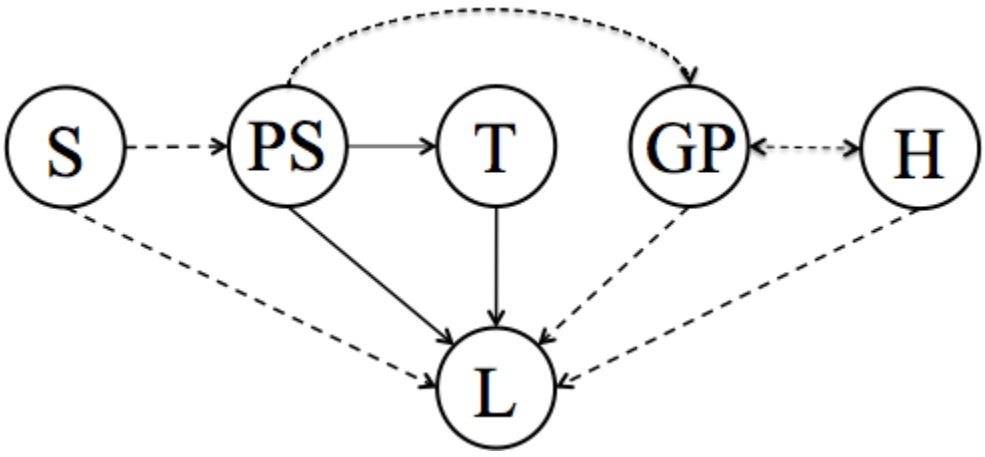
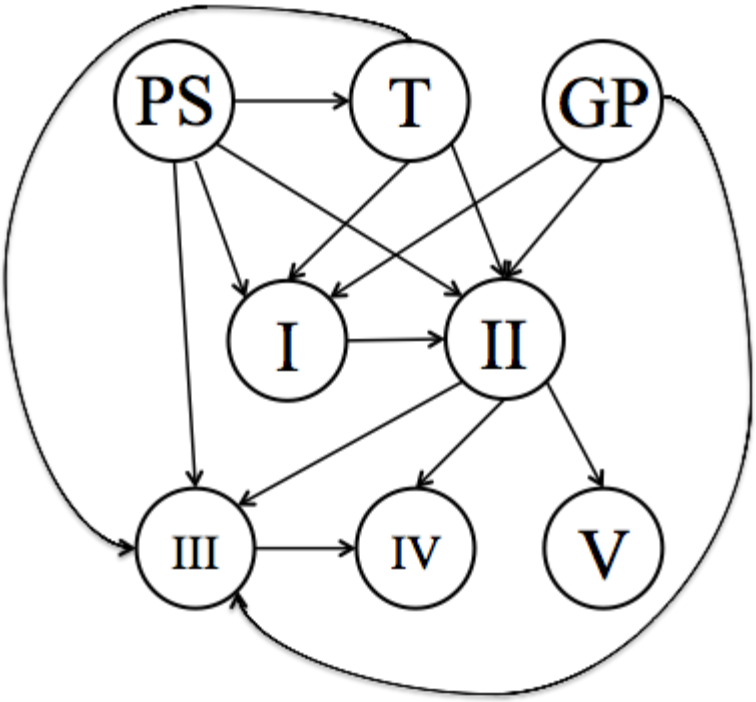


Figure 2. Bayesian network model structure for the evaluation model. Node PS represents primary site of the tumor; node T represents T-stage/size of the tumor, node GP represents genetic profile of the patient, node H represent HPV status of the patient, node S represents the smoking history indicator, node L represent surgical levels of the neck. The directed edges (arrows) indicate conditional relationships between variables. The variables at arrowheads are conditionally dependent on the variables at the tails of the arrows. The dotted edges are relationships to be evaluated.

2.1.2 *Prototype Model Structure*

After evaluating the effects of additional predictors on metastasis, we developed a prototype model that accounted for the anatomical lymphatic structure of the neck, as well as all skip lesions and branching scenarios, represented by the conditional dependencies between different levels of metastasis (Figure 3). The outcome of each level is binary (positive or negative), and can be a predictor for other levels. Only genetic profile information was added as the additional predictor based on the results from evaluation model. Also, genetic profile only affects the outcomes of the levels that HNSCC can initially metastasizes to, which are level I, II



and III.

Figure 3. Bayesian network model structure for the prototype model. Node PS represents primary site of the tumor; node T represents T-stage/size of the tumor, node GP represents genetic profile of the patient, node I through V represent surgical levels of the neck. The directed edges (arrows) indicate conditional relationships between variables. The variables at arrowheads are conditionally dependent on the variables at the tails of the arrows.

2.2 Data Sources

We used two data sources. The first data source is from University of Washington tumor board (UW TB). We obtained records of 383 patients with untreated, non-recurrent squamous cell carcinoma (SCCA) of the head and neck presented to the UW head and neck tumor board over a 3.5-year period. Since this dataset does not contain either patient genetic information or HPV status, we used this for training the baseline model.

The second data source is from the Cancer Genome Atlas (TCGA)[14]. We exported 528 subjects' clinical data, pathology reports, HTSeq-FPKM-UQ (upper quintile of normalized gene

expression values) and HTSeq-Counts (raw gene expression values) from the TCGA data portal.

2.3 TCGA Data Preparation

2.3.1 *Clinical Data*

Unique identifiers, primary site information, clinical T stage, HPV status, smoking history indicator were extracted from subjects' clinical data. Because of the limited sample size and the relatively wide range of different primary sites, the number of subjects for each primary site is too small to be statistically significant. Hence, the primary sites were aggregated into larger regional sites as follows: "Tongue", "Floor Of Mouth", "Oral Commissure", "Lip". "Alveolar Ridge", "Buccal Mucosa" and "Mandible" are represented as "Oral Cavity"; "Base Of Tongue", "Tonsil", "Retromolar Trigone" are represented as "Oropharynx"; "Supraglottic larynx" and "Glottic Larynx" are represented as "Larynx". Some primary sites (e.g. "Palate") contained too few samples and were not included. Subjects with unknown T-stage were dropped. Those with "T4a" and "T4b" T-stages were combined into stage "T4".

Each patient's level of nodal metastatic information was manually extracted from pathology reports, linked to corresponding subjects based on the TCGA manifest, and reviewed by two team members for accuracy. Only 349 out of 528 subjects had data on their nodal metastatic level. The data are presented as the number of positive lymph nodes in each level. For the evaluation model (Figure 2), we only used information on the highest level of nodal metastasis, for example, if both level I and level III have positive nodes, level III is reported for this subject as the highest level of metastasis. For the prototype model (Figure 3), the nodal metastasis level data were turned into boolean variables (positive/negative) for each level.

2.3.2 *Gene Expression Data*

Since studies performed by Mendez et al. have identified the tumor-specific genes

differentially expressed between metastatic and non-metastatic oral cancer [8][9], we applied this established gene expression profile to the TCGA dataset. Because only oral cancer patients were studied by Mendez et al. and our dataset has other types of HNSCC, we first validated the gene expression profile using the TCGA dataset. The HTSeq-Counts of the subjects in this dataset were analyzed to identify differentially expressed genes between patients who had lymphatic metastasis and those who did not. The differential analysis was performed using the R Bioconductor package, version 3.4, DESeq2. The resulting genes were ranked based on p-value and used to cross-reference with the tumor-specific genes determined in Mendez et al.'s studies [8][9]. Of the genes we identified from the TCGA dataset, 53 genes match to genes discovered in the previous Mendez studies. All 53 matched genes were associated with low p-values that fall within the first half of the ranked gene list, which supported our decision to apply this established gene expression profile to our dataset.

Because the 53 genes are not equally predictive of the level that the lymph nodes would be affected, having all of them as individual predictors will be likely to include the low-effect ones. We first applied a Principle Component Analysis (PCA) and a Random Forest model aiming to reduce the dimensionality. However, results from both dimension reduction methods required us to set an arbitrary threshold of inclusion (supplement figure 1 & 2). Therefore, we decided to use an aggregation approach and applied the multivariate logistic regression model published in Dr. Mendez's studies to fit our HTSeq-FPKM-UQ data and the metastatic status [8][9].

The log odd calculated from fitting the regression model were used as propensity score for each patient to indicate the risk of metastasis associated with his or her genetic profile. The normalized propensity scores of the 349 subjects have the following distribution (Figure 4). They

were then categorized based on effect sizes with 95% confidence intervals to help clinicians to identify patients who are more prone to develop metastasis. Specifically, subjects with high propensity scores are classified into "high" risk category (n = 56), subjects with low propensity scores are classified into "low" risk category (n = 54), and the rest of the subjects fall under the "regular" risk category (n = 239).

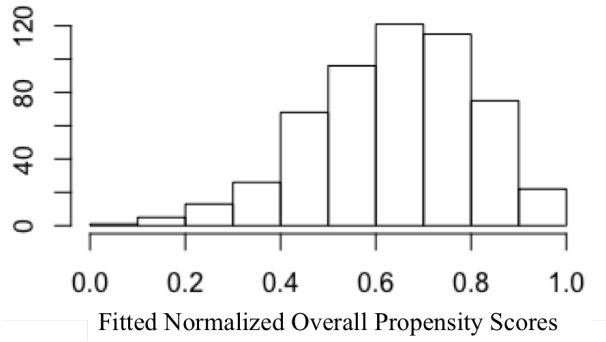


Figure 4. Histogram of normalized overall propensity scores of the 349 subjects

2.4 Parameter Learning/Conditional Probability Table

From the two data sources, TCGA and UW tumor board, we produced datasets described as follows:

1. TCGA data with genetic expression profile, sample size = 349, training set = 233
2. TCGA data with smoking history indicator, sample size = 349, training set = 233
3. TCGA data without any additional predictors, sample size = 349, training set = 233
4. UW tumor board data without any additional predictors, sample size = 383, training set = 255
5. Combination of TCGA and UW tumor board data without any additional predictors, sample size = 595, training set = 392. When combining the datasets, we tried to preserve the primary site distribution of the TCGA dataset by proportionally adding data from the UW tumor board dataset, data points were randomly selected but reproducible.

The evaluation model was trained and tested using the R "bnlearn" package (Version 4.1.1) with the 3-fold cross-validation method with each of the above dataset. Conditional probability tables were produced from fitting the training set; predictions were made on their corresponding test sets. The model performance and prediction accuracies were measured and compared by calculating the correlation coefficients, cosine similarities and the Area Under the Curve (AUC) values. Because there is no significance test for comparing model accuracies, we used these three measures to ensure the validity of our results. The correlation coefficient measures the model performance by correlating the predicted probability of each distinct combination of predictors learned from the training set with the observed probability from the test set. Cosine similarity measures prediction accuracy by calculating the cosine of the angle between the predicted level of nodal metastasis and the observed one in the test set. The weighted AUC measures discrimination, that is, the ability of the model to correctly predict subjects' highest metastatic level. Since our model output has more than two classes, we generated ROC curves for each outcome independently and then averaged them while taking into account the sizes of the classes.

The prototype model was trained and tested with the TCGA dataset with and without genetic profile, UW tumor board dataset and the combination dataset. Since there are five outcome variables, one for each level, prediction accuracies were only reported for level III as an example. The prediction accuracies were compared with those from training the evaluation model, aiming to investigate the effects of incorporating anatomical lymphatic structure on model performance and prediction accuracies.

2.5 Decision Support Table and Application

The estimated probabilities of having positive lymph nodes in each level were

categorized into “treat,” “maybe treat”, and “do not treat” based on a consensus reached by physicians (< 7% = do not treat, between 7% and 15% inclusive = maybe treat, > 15% = treat) as reported in a previous study [15].

For the evaluation model, we studied the effect of adding genetic information to these decision support tables, and whether it improves clinically relevant decision making. The decision support tables were only constructed for subjects with T4 “Oral Cavity” tumors because data on these subjects is most complete amongst the rest and the decision support tables have the least number of missing values.

For the prototype model, we fitted the entire TCGA dataset to produce the conditional probability tables, which were used to build the decision support tables for each combination of the independent variables. A Shiny Application was developed to allow users to select different features of the patient and get the resulting decision support table.

CHAPTER 3 RESULTS

3.1 EVALUATION MODEL

The distributions of primary sites, T-stage, and positive nodal metastasis of patients in the training datasets are given in Table 1 - 3.

Table 1– Datasets Comparison – Primary Sites

Primary Sites	**TCGA n = 349**	**UW Tumor Board n = 383**	**Combination n = 595**
Oral Cavity	0.70	0.37	0.65
Larynx	0.19	0.20	0.20
Oropharynx	0.10	0.42	0.14

Table 2– Datasets Comparison – Clinical T Stage

T Stages	**TCGA n = 349**	**UW Tumor Board n = 383**	**Combination n = 595**
T1	0.07	0.18	0.13
T2	0.27	0.34	0.30
T3	0.27	0.19	0.24

T4	0.37	0.28	0.31

Table 3– Datasets Comparison – Highest Nodal Metastasis Level

Nodal Metastasis Level	**TCGA n = 349**	**UW Tumor Board n = 383**	**Combination n = 595**
No Met	0.51	0.33	0.46
Level I	0.09	0.08	0.08
Level II	0.16	0.28	0.20
Level III	0.13	0.20	0.14
Level IV	0.08	0.08	0.07
Level V	0.03	0.03	0.03

The cosine similarities of different data sets (Table 4) indicate some heterogeneity between TCGA and UW tumor board data. Specifically, TCGA dataset contains a larger proportion of subjects with "Oral Cavity" tumors (70% in TCGA; 37% in UW TB), whereas the UW tumor board dataset has a larger proportion (10% in TCGA; 42% in UW TB) of subjects with "Oropharynx" tumor (Table 1). The difference in distribution may affect our results for prediction accuracy comparison between models trained with different datasets. This is because for the evaluation model, we made a simplified assumption that metastasis follows a linear fashion, which is more representative of the metastasis of oral cavity cancer. For the UW tumor

board dataset, which has more other types of head and neck cancers, the prediction accuracy could be worse and make our results difficult to interpret. Therefore, when combining the two datasets, the primary site distribution of TCGA dataset was maximally preserved.

Table 4– Cosine similarities between the datasets

	TCGA vs. UW TB	**UW TB vs. Combination**	**TCGA vs. Combination**
Cosine similarity	0.86	0.90	0.99

3.1.1 *Sample Size And Model Performance*

The comparison of baseline model prediction accuracies learned from the two different data sources with different sample sizes is shown in Table 5. None of these models incorporated additional predictors.

Table 5– Model performance and prediction accuracy comparison - two data sources

Dataset	**Training Set Sample Size**	**Correlation Coefficient**	**Cosine Similarity**	**AUC**
TCGA	233	0.70	0.66	0.65
UW Tumor Board	255	0.63	0.70	0.57
Combination	392	0.86	0.77	0.72

The correlation coefficient, cosine similarity and weighted AUC of the model trained

with the combination dataset are the highest amongst all the groups. It means the model has relatively better model performance, the differences between the predictions and the observations are relatively small and that this model resulted in the least amount of classification errors. The improvement in prediction accuracies is most likely due to the increase in sample size, as the combination dataset has the most amounts of data.

The model trained with the UW Tumor Board dataset has relatively poor prediction accuracies, although its sample size is a little bit larger than that of the TCGA dataset. The reason could be that not accounting for the lymphatic structure and dependencies between metastatic levels is a poorer assumption to make for tumors developed from certain primary sites. The UW Tumor Board dataset contains more subjects with Oropharynx cancer and less Oral Cavity cancer comparing to the TCGA dataset (Table 1), and Oropharynx cancer actually does not metastasize to level I or level V of the neck.

3.1.2 *Genetic Profile And Model Performance*

Table 5 shows the comparison between the prediction accuracies trained with baseline model and the ones with additional predictors. All models were trained and tested with the TCGA dataset.

Table 5– Model performance and prediction accuracy comparison – TCGA baseline model with additional predictors

Additional Predictors	Correlation Coefficient	Cosine Similarity	AUC
None	0.70	0.66	0.65
Genetic Profile	0.80	0.74	0.71
Smoking	0.51	0.67	0.60

The model with genetic profile has higher correlation coefficient, cosine similarity and AUC comparing to the baseline model, indicating that adding this feature can improve model performance and prediction accuracy. The model with smoking as the additional predictor has lower measures of accuracy compared to the baseline model, which supports our hypothesis that smoking does not affect metastasis although it's a strong predictor for the onset of HNSCC.

3.1.3 *Genetic Profile And Probability of Metastasis*

Table 6 shows the comparison of the estimated probabilities of a subject with T4 "Oral Cavity" cancer having metastasis to specific nodal levels, as predicted by the models with and without genetic profile information as a predictor. Figures 5,6 are the visualizations for the probabilities in Table 6.

Table 6– Conditional Probability Table Comparison for T4 Oral Cavity tumor metastasis

Model	TCGA All	TCGA High Risk	TCGA Regular Risk	TCGA Low Risk	UW Tumor Board
No mets	0.43	0.21	0.48	0.43	0.33
Level I	0.16	0.20	0.15	0.15	0.18
Level II	0.21	0.39	0.18	0.21	0.27
Level III	0.08	0.10	0.05	0.15	0.09
Level IV	0.08	0.00	0.10	0.05	0.12
Level V	0.04	0.10	0.05	0.00	0.00

TCGA whole: TCGA data without genetic expression profile (n = 233). TCGA High: TCGA data with high propensity score (n = 40). TCGA Regular: TCGA with regular propensity score (n = 155). TCGA Low: TCGA with low propensity score (n = 38) UW TB: UW Tumor Board (n = 255)

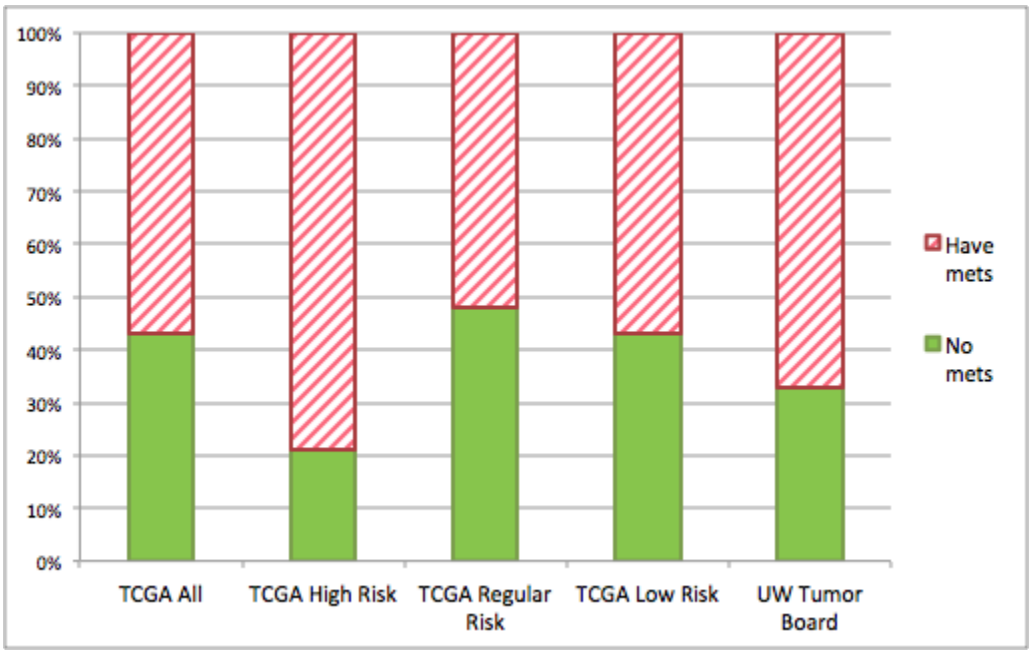


Figure 5. Comparing the effects of low, regular, and high-risk genetic profile on metastasis

status. Red area represents proportions of subjects in datasets with metastasis; Green area represents proportions of subjects in datasets without metastasis.

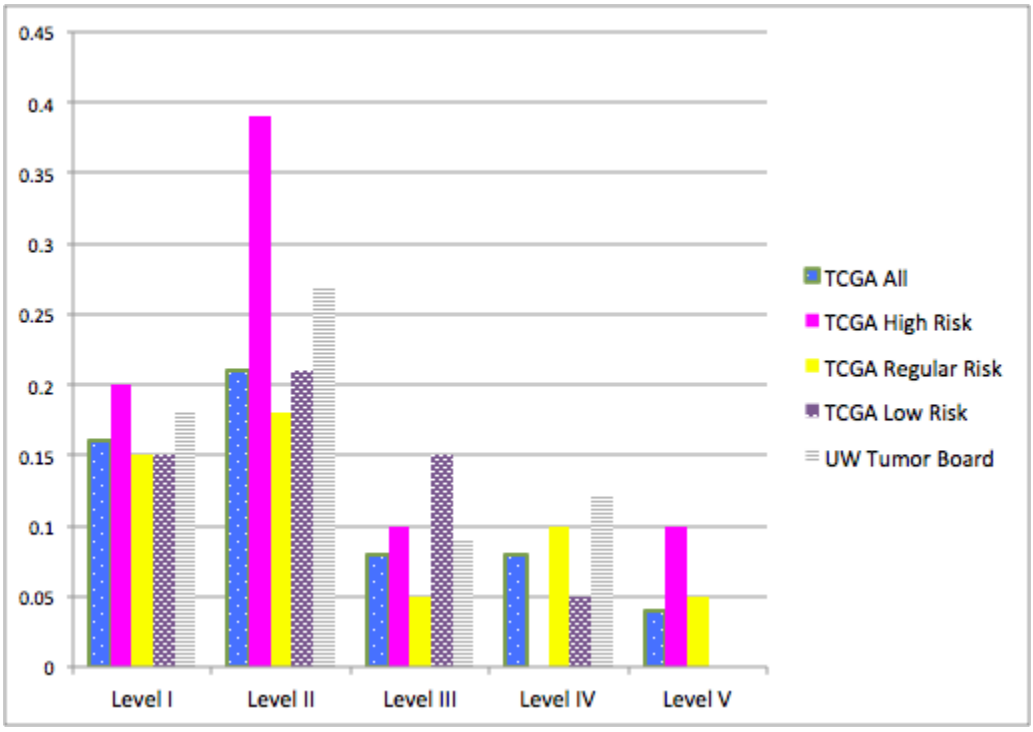


Figure 6. Comparing the effects of low, regular, and high-risk genetic profile on metastasis to certain levels. Blue bars represent subjects from the entire TCGA dataset: Pink bar represents subjects from the TCGA dataset with a High risk genetic profile. Yellow bars represent subjects from the TCGA dataset with a Regular risk genetic profile. Purple bars represent subjects from the TCGA dataset with a Low risk genetic profile. Grey bars represent subjects from the UW Tumor Board dataset.

Probabilities shown in Table 6 and Figure 5 imply that patients with high propensity scores are more likely to develop nodal metastasis regardless of levels (79% of the subjects with "high" risk had metastasis; 57% of the subjects with "low" risk had metastasis. They are also more prone to have level I and II nodal metastasis (pink bar in Figure 6). This seems valid because Oral Cavity tumor initially metastasizes to level II and level I (level II more often than

level I), and the genetic profile is predictive of the onset of metastasis rather than the prognosis. However, there is not enough evidence to show significance of these findings because of the relatively small sample sizes, especially for the groups with high and low propensity score.

Table 7– Probability with Treatment Decision Comparison for T4 Oral Cavity tumor metastasis

Model	UW Tumor Board	TCGA All	TCGA High Risk	TCGA Regular Risk	TCGA Low Risk
No mets	0.33	0.43	0.21	0.48	0.43
Level I	0.66	0.57	0.79	0.52	0.57
Level II	0.48	0.41	0.59	0.38	0.41
Level III	0.21	0.20	0.20	0.20	0.20
Level IV	0.12	0.12	0.10	0.15	0.05
Level V	0.00	0.04	0.10	0.05	0.00

Shading of the cells indicates treatment decisions. Red represents "treat"; yellow represents "maybe treat"; green represents "do not treat".

Table 7 shows the comparison of the estimated probability of finding positive lymph nodes in each level for subjects with T4 "Oral Cavity" cancer and the treatment decision based on decision thresholds. The treatment decisions from training the entire TCGA dataset and the UW tumor board datasets are the same (leftmost two columns in Table 7), but they are different for subjects with different genetic profiles. Comparing to subjects with "regular" risk genetic profiles and subjects from datasets without genetic information, the ones with "high" risk genetic

profiles show a change in decision at level V (from “do not treat” to “maybe treat”); subjects with “low” risk genetic profiles show a change in decision at level IV (from “maybe treat” to “do not treat”). This result seems reasonable, as more treatments should be considered for patients with a high risk of developing metastasis.

3.2 PROTOTYPE MODEL

Even though the intention for constructing the prototype model was to build the decision support application, given the available results on the evaluation model, we could easily study the effect of incorporating the anatomical structure of the lymphatic system in the neck by comparing the two models’ performances. Table 9 compares the prediction accuracies and model performance for level III nodal metastasis from training the prototype model and the evaluation model using the four different datasets.

Table 9– Model performance and prediction accuracy comparison – Level III different model structure and different datasets

Dataset	Model Structure	Correlation Coefficient	Cosine Similarity	AUC
TCGA with genetic profile information	Prototype Model	0.83	0.95	0.61
TCGA with genetic profile information	Evaluation Model	0.80	0.74	0.71
TCGA without genetic profile information	Prototype Model	0.73	0.95	0.52
TCGA without genetic profile information	Evaluation Model	0.70	0.66	0.65
UW Tumor Board	Prototype Model	0.72	0.98	0.57
UW Tumor Board	Evaluation Model	0.63	0.70	0.57
Combination	Prototype Model	0.87	0.95	0.61
Combination	Evaluation Model	0.86	0.77	0.72

The prototype model has better model performance, reflected by the relatively higher correlation coefficients from all four datasets. This means that incorporating the lymphatic anatomical structure into the model improved its representativeness of the data, however, this also means that the model may have been overfitted. In terms of cosine similarities, because the outputs for the prototype model are binary (negative or positive; 0 or 1) whereas the evaluation model's output has five possibilities (level 0 – 5), they are not comparable. The prediction

accuracies, AUC values, of the prototype model are all worse than those of the evaluation model. This could be explained by the significantly smaller training set sample size now that the nodal metastasis in some levels are dependent on the others.

Comparing the prediction accuracies between prototype model trained with different datasets (rows with grey background in Table 9), the one with genetic profile information and the one trained with the most data still have the highest prediction accuracies amongst the rest, which agrees with the comparison results from training the evaluation model.

The resulting conditional probability tables from fitting the entire TCGA dataset with genetic information into the prototype model are built into a decision support application shown in figure 7. The web Application is available at: https://sw21.shinyapps.io/met_pred/. Users can select different primary site, T stage, risk associated with patient's genetic profile, and metastasis status of level I, II and III on the side panel, click "show predictions" button, and get the probabilities of having metastasis in each level of the neck, along with the medical decision suggestions. A figure of the lymphatic drainage pathway for the primary site selected will also be shown on the bottom of the main panel for better clarity.

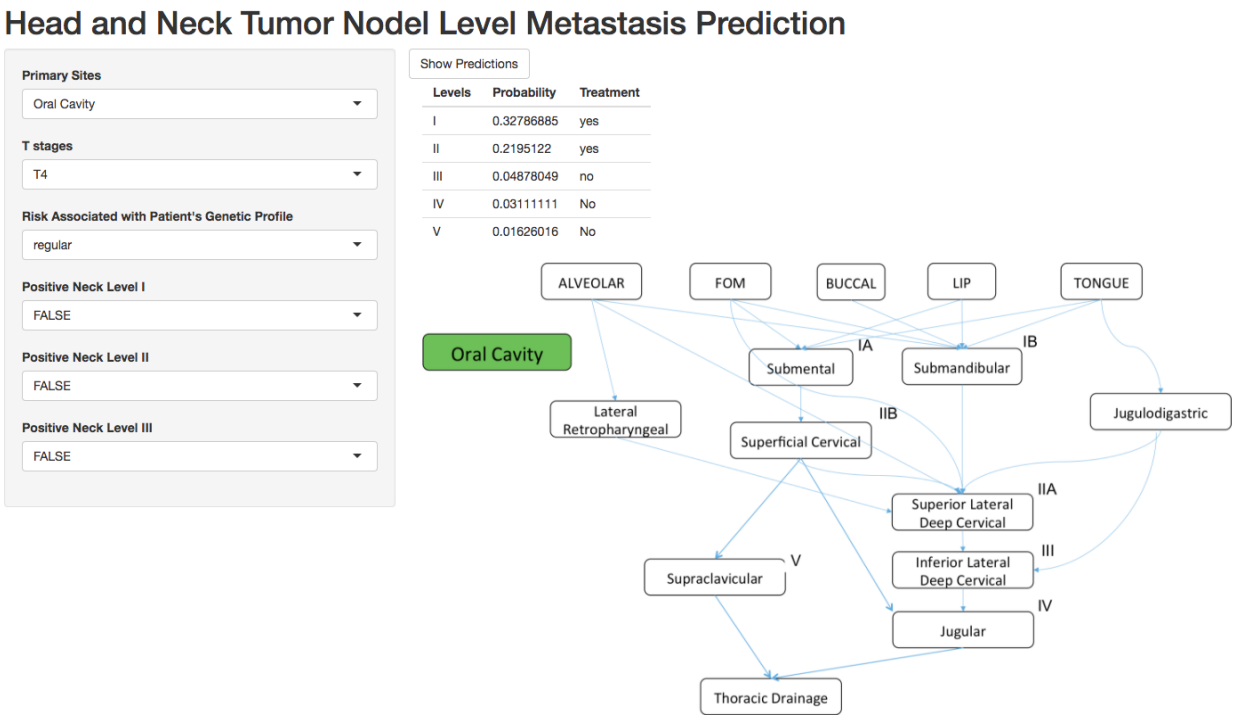


Figure 7: Interface of the Decision Support Application Prototype

CHAPTER 4 DISCUSSION

Based on the results from both the evaluation model and the prototype model, all three accuracy measures (correlation, cosine similarity and AUC) showed evidence that having a genetic profile as an additional predictor improved model's prediction accuracies compared to models trained without it. The improvements seem to be equal to that of having a larger training set when we compared the prediction accuracies between TCGA with genetic profile (n = 233) and the Combination (n = 392) dataset (rows with grey background in Table 9). This is promising, since it is often more difficult to recruit a larger sample population than to examine all possible predictors in a relatively smaller sample.

Because we used the genetic profile identified in the Mendez et al. [9][10] studies and the associated parameters while preparing the training sets for our model, our results further proved

the validity of using the genetic profile to predict metastasis in HNSCC patients. It remains to be investigated whether the genetic profile predictive of metastasis can also be predictive once the cancer cells have spread into the lymphatic system. If there exists a genetic profile associated with more aggressive cancer cells leading to a higher level of metastasis, it could provide a future research direction.

Due to the limited sample size, the current prototype model structure reflected all variations and possibilities in metastatic pathways for all of the primary sites because we aggregated the primary sites into the three regions (Oral Cavity, Oropharynx, Larynx). Since primary sites have differences in metastatic potential even within a region, poorer performance is expected from aggregating into regions than if we had enough data to model individual primary sites. Ideally, models should be constructed separately to represent the metastasis pathway for each individual primary site to allow precision and easy interpretation of the results. In addition, when constructing the models, structure learning from the data should be able to provide validation for the actual relationships between each level of the neck. However in our case, because many of the combinations of predictor variables only corresponded to either a few or none of the subjects, the prototype model structure was constructed based on the lymphatic anatomical structure, and structure learning could not provide much insight. For the same reason, the prediction accuracies of the prototype model were lower than those of the evaluation model and the results from training this model should only be used to demonstrate the ability of the decision support tool and not be used in clinical settings for now.

The major limitation of this study is the small sample sizes. Because not all pathology reports include the recording of metastatic levels, it was difficult for us to gather this data. Data including patient genetic information is even more sparse, as only certain research studies

currently collect the patient's genetic information for analysis. We therefore needed to use cross-validation rather than a separate dataset with both genetic profile information and metastatic level information to validate our model. Moreover, even though we were able to quantify the effects of genetic profile on metastasis, we could not show evidence that such effects are statistically significant and can lead to changes in treatment decisions. However, because the treatment decisions do not change much depending on the dataset used to train the model (Table 8), it means that our baseline model is consistent in terms of treatment granularity across all data. As the TCGA dataset being continuously updated, perhaps more data will become available in the future for us to improve the methods and results of this project to achieve better model performance and be able to make more significant conclusions.

Despite the limitations, our study is the first to quantify the effects of incorporating additional predictor, such as genetic profiling, on the level of metastasis in HNSCC patients. It is also an indication that genetic information can be used to assist medical decision-making in this domain, although the effect size is inconclusive. This calls for an increased collection of relevant data to determine the significance.

CHAPTER 5 CONCLUSION

Our predictive model depicted and quantified the relationships between the metastatic level in patients with HNSCC and potential predictors such as tumor primary site, T-stage, and metastatic risk propensity derived from patient genetic profiles. The results supported our hypothesis that gene expression have effects on metastasis, and that including it as a predictor can improve model accuracy. Although there was not enough data to show the significance of the effects that different genetic profiles have on treatment decision-making, a decision support

application was developed to demonstrate the clinical significance of our model. Overall, our predictive model can improve scientific knowledge and clinical practice in HNSCC treatment and can be further improved with more data.

REFERENCES

[1] R.L. Siegel, K.D. Miller, and A. Jemal, Cancer statistics, 2016, *CA Cancer J Clin* **66** (2016), 7-30.

[2] R.J. Sanderson and J.A.D. Ironside, Squamous cell carcinomas of the head and neck, *Bmj* **325** (2002), 822-827.

[3] M.J. Ruback, A.L. Galbiatti, L.M. Arantes, G.H. Marucci, A. Russo, M.T. Ruiz-Cintra, L.S. Raposo, J.V. Maniglia, E.C. Pavarino, and E.M. Goloni-Bertollo, Clinical and epidemiological characteristics of patients in the head and neck surgery department of a university hospital, *Sao Paulo Med J* **130** (2012), 307-313.

[4] P.M. Som, H.D. Curtin, and A.A. Mancuso, Imaging-based nodal classification for evaluation of neck metastatic adenopathy, *AJR Am J Roentgenol* **174** (2000), 837-844.

[5] K.T. Robbins, J.E. Medina, G.T. Wolfe, P.A. Levine, R.B. Sessions, and C.W. Pruet, Standardizing neck dissection terminology. Official report of the Academy's Committee for Head and Neck Surgery and Oncology, *Arch Otolaryngol Head Neck Surg* **117** (1991), 601-605.

[6] N. Benson, M. Whipple, and I.J. Kalet, A Markov Model Approach to Predicting Regional Tumor Spread in the Lymphatic System of the Head and Neck, *AMIA Annu Symp Proc* **2006** (2006), 31-35.

[7] P. Croskerry, From mindless to mindful practice--cognitive bias and clinical decision making, *N Engl J Med* **368** (2013), 2445-2448.

[8] A. Law, S. Wu, H. Jung, L. Wang, E. Grunblatt, M. Whipple, Prediction of regional

metastasis in squamous cell carcinoma of the oral cavity, In preparation.

[9] E. Mendez, W. Fan, P. Choi, S.N. Agoff, M. Whipple, D.G. Farwell, N.D. Futran, E.A. Weymuller, Jr., L.P. Zhao, and C. Chen, Tumor-specific genetic expression profile of metastatic oral squamous cell carcinoma, *Head Neck* **29** (2007), 803-814.

[10] E. Mendez, P. Lohavanichbutr, W. Fan, J.R. Houck, T.C. Rue, D.R. Doody, N.D. Futran, M.P. Upton, B. Yueh, L.P. Zhao, S.M. Schwartz, and C. Chen, Can a metastatic gene expression profile outperform tumor size as a predictor of occult lymph node metastasis in oral cancer patients?, *Clin Cancer Res* **17** (2011), 2466-2473.

[11] A.L.S. Galbiatti, J.A. Padovani-Junior, J.V. Maníglia, C.D.S. Rodrigues, É.C. Pavarino, E.M. Goloni-Bertollo, Head and neck cancer: causes, prevention and treatment, *Brazilian Journal of Otorhinolaryngology*, Volume 79, Issue 2, 2013, Pages 239-247, ISSN 1808-8694,

[12] N. Benson, M. Whipple , I.J. Kalet. A Markov Model Approach to Predicting Regional Tumor Spread in the Lymphatic System of the Head and Neck. *AMIA Annual Symposium Proceedings.* 2006;2006:31-35.

[13] C. Rosse, L.G. Shapiro, and J.F. Brinkley, The digital anatomist foundational model: principles for defining and structuring its concept domain, *Proc AMIA Symp* (1998), 820-824.

[14]Cancer Genome Atlas (TCGA). Genomic Data Commons Data Portal: https://gdc-portal.nci.nih.gov/

[15]I.J. Kalet, M. Whipple, S. Pessah, J. Barker, M.M. Austin-Seymour, and L.G. Shapiro, A rule-based model for local and regional tumor spread, *Proc AMIA Symp* (2002), 360-364.