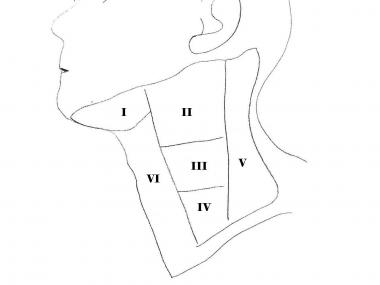
**Final Project**

Shuyang Wu

**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 locations of lymphatic metastasis is critical for both minimizing the risk of recurrence and minimizing the complications resulting from unnecessary radiation.

*Figure 1 (left). 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.*

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.

Studies have shown that there is a genetic expression profile predictive of nodal metastasis of HNSCC [8][9]. Other risk factors such as smoking, alcohol and human papillomavirus (HPV) status [10] are also shown to have effects on metastasis. Therefore, we propose to develop a predictive model to capture the probabilities of finding cancerous lymph nodes in each neck level using a patient’s genetic profile, and clinical information including primary tumor site, tumor size, hpv status. Such a model could potentially help physicians to make improved evidence-based decisions while performing targeted treatments such as radiation therapy.

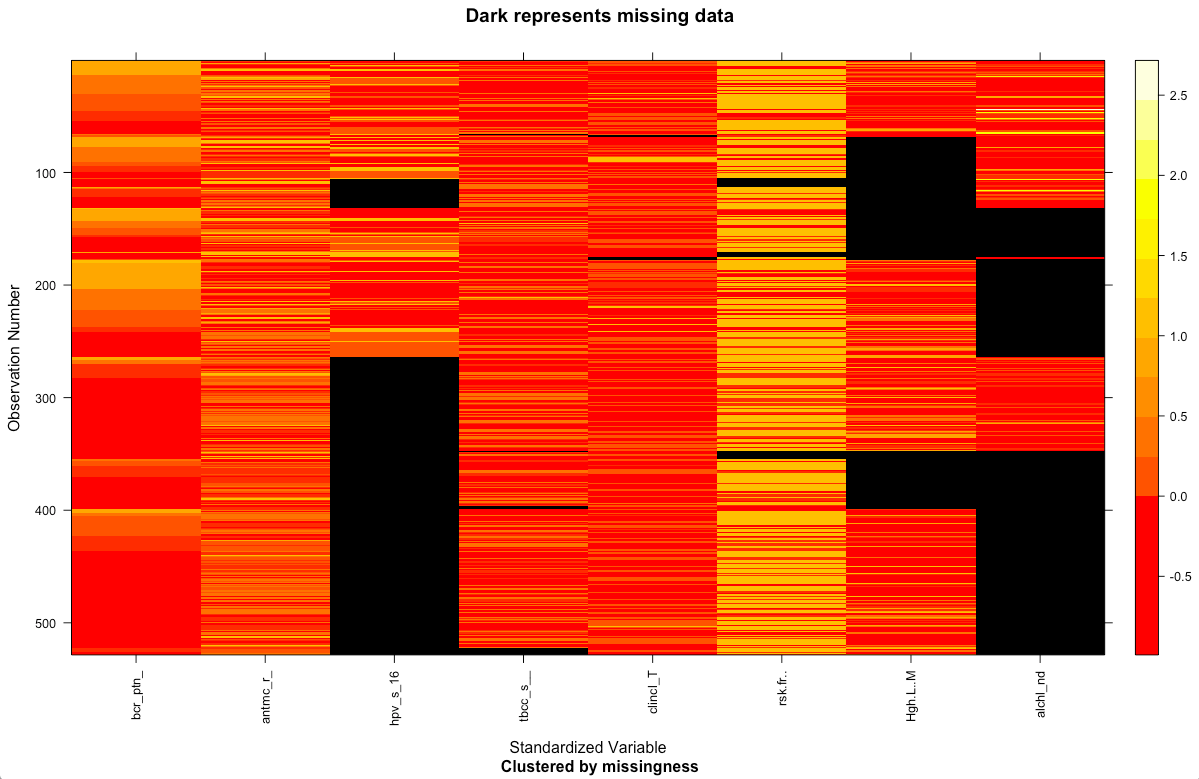
**About the Data**

All data were extracted from the Cancer Genome Atlas. Some raw data has been cleaned, for example, the Alcohol consumption frequency and risk from genetic profile were turned into categorical indicators. Table 1 shows each variable’s name, dependency, data type and missingness. Figure 2 shows the missing pattern for each variable.

Table 1. Summary for each variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Dependent/Independent | Data Type | Number of Missing Entries | Missing Pattern |
| Primary Site | Independent | Unordered Categorical  3 levels | 0/528 | N/A |
| T stage | Independent | Ordered Categorical  5 levels | 4/528 | MCAR |
| Risk from Genetic Profile | Independent | Ordered Categorical  3 levels | 20/528 | MCAR |
| Smoking history indicator | Independent | Ordered Categorical  5 levels | 10/528 | MAR |
| Alcohol consumption indicator | Independent | Ordered Categorical  5 levels | 307/528 | MAR |
| HPV | Independent | Boolean | 291/528 | MCAR |
| Highest Level of Metastasis | Dependent | Ordered Categorical  6 levels | 139/528 | N/A |

Figure 2. Missing pattern for each variable



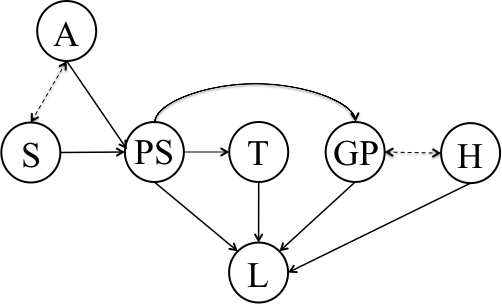
Patient ID Primary Site HPV Smoking T stage Risk from Highest level Alcohol

genetic profile of metastasis

The missing entries for “Highest level of metastasis” will have to be dropped because it’s the dependent variable, reducing the sample size to 368. For the other variables with missing entries, since there is no way to track down how data was collected, it’s difficult to determine if any of them are MNAR. By modeling each variable’s missingness with variables’ values, I was able to determine whether they are MCAR or MAR. Although most of the variable’s missingness is MCAR, which is ignorable, dropping all subjects with a missing entry would significantly reduce the sample size. Thus, missing data should be imputed for modeling.

**Proposed Model**

Based on previous research and expert knowledge, we know that some of these variables are dependent on others. Bayesian networks works well in representing complex stochastic relationships between interacting variables, therefore is the model I will use. A Bayesian network is a Directed Acyclic Graph (DAG) with the directed edges representing the conditional independencies between the variables and parameters representing the conditional probabilities or distribution of the outcome variables. The structure of the Bayesian Network is often learned from the data, but as we have sufficient domain knowledge in this area, this model’s structured is pre-specified, as shown in figure 2. Data will be used to validate the structure.

*Figure 2. Bayesian network model structure. Node PS represents primary site of the tumor; node T represents T-stage/size of the tumor; node GP represents risk associated with the genetic profile of the patient; Node A represent patients’ alcohol consumption index; node S represent smoking frequency index; node H represent hpv status; node L represents highest level of tumor metastasis. 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. Dashed lines indicate relationships that remain to be validated with data.*

**Missing Data Imputation**

Both EM and Bayesian Inference are popular methods to infer missing values in Bayesian Network models. I would like to use EM algorithm in my model. The assumptions involved in EM algorithms are: parametric estimation; the missing data mechanism is ignorable, missing data pattern is MAR; and the parameters for missingness are distinctive of the parameters for the observations. For my model, given the limited information I have on how data was collected, I made the assumption that all data was MAR for now (to simplify the likelihood function).

In EM, random parameters will be put in the model to impute the missing entries, complete case analyses will be done and conditional expectation of the complete-data log-likelihood will be calculated. And then the parameters will be updates as log-likelihood being maximized (until converges).

One advantage of using EM in my model is it works easily with categorical data, which aligns with the data types of all variables in my dataset. It’s relatively easy to program and fairly efficient. It will converge if it can, but it’s possible to stop at local minima as my model has a significant number of features. Thus, I will try with many different starting inputs to try to avoid such problem. A disadvantage is that the process might be slow because I have a fairly large fraction of missing entries. Also, since it only uses first derivative when maximizing, it might be computationally inefficient. In addition, the calculation of standard error and confidence intervals will involve bootstrapping, which is not very straightforward.

**Proposed Methods**

The dataset will be randomly split into training set and test set K times for the purpose of K-fold cross-validation. Because the dependent variable has six levels, six conditional probability tables from parameter learning will be produced using the R “bnlearn” package (Version 4.0). The trained models will then be used to predict their corresponding test sets. The model prediction accuracies will be measured and reported. One possible measure of accuracies is the Area Under the Curve (AUC) values. The AUC measures discrimination, which is the ability of the model to correctly predict subjects’ highest metastatic level. Since my model output has more than two classes, I will need to generate ROC curves for each possible outcome independently and then average them while taking into account the sizes of the class.

The estimated probabilities of having positive lymph nodes in each level will be 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 [11]. Decision support tables will be constructed for subject with different combinations of features. Eventually, the results of this model will have an interactive interface for physicians to reference.

**Reference**

[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] 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.

[9] 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.

[10] R. Sankaranarayanan, E. Masuyer, R. Swaminathan, J. Ferlay, and S. Whelan, Head and neck cancer: a global perspective on epidemiology and prognosis, Anticancer Res 18 (1998), 4779-4786.

[11]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.