Spring 2014

Outcome prediction in head and neck cancer patients using machine learning methods

David John Dellsperger
University of Iowa

OUTCOME PREDICTION IN HEAD AND NECK CANCER PATIENTS USING MACHINE LEARNING METHODS

by

David John Dellsperger

A thesis submitted in partial fulfillment
of the requirements for the Master of
Science degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

May 2014

Thesis Supervisor: Professor Thomas L. Casavant

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

David John Dellsperger

has been approved by the Examining Committee
for the thesis requirement for the Master of Science
degree in Biomedical Engineering at the May 2014 graduation.

Thesis Committee: _____
Thomas L. Casavant, Thesis Supervisor

_____
Terry Braun

_____
Todd Scheetz

This thesis is dedicated to my Parents Kevin and Linda, Brother Ben and wife Catheryn, Grandparents John and Clare Martinkovic, and Doris and Charles Dellsperger, Aunts and Uncles, Cousins, and my friends who have helped me to become the person I am today.

ABSTRACT

Head and Neck cancers account for approximately 3.2% of the estimated 1,660,290 new cancer cases for the year 2013 and roughly 1.9% of cancer-related deaths in 2013. In this research, machine learning techniques were employed to predict outcome in cancer patients supporting more objective assessment of the treatments, including surgery, radiation therapy, or chemotherapy. Selection of features capable of distinguishing between the possible outcomes was accomplished by using a highly selective cohort of 61 patients with similar treatment and location of the primary tumor. An accuracy of 80.33% (compared to a baseline majority classifier of 60.66%) was achieved utilizing this cohort. Further, it is shown that this limited cohort has the power to provide valuable information on outcome prediction utilizing as few as four features. Feature selection was drawn from both clinical features and quantitative imaging features including the site of cancer, primary tumor volume, and race.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Head and Neck cancers account for approximately 3.2% of the estimated 1,660,290 new cancer cases for the year 2013 and roughly 1.9% of cancer-related deaths in 2013. (American Cancer Society, 2013) Some of the common treatments for head and neck cancer include Surgery, Radiation Therapy, Chemotherapy, or a combination of these treatments. (National Cancer Institute, 2012) The University of Iowa is one of the 17 current members of the Quantitative Imaging Network (QIN) (National Institute of Health, 2013) with the goal of discovering methods for utilizing image-based quantification of response to cancer patient treatments. The practical objective of this large-scale effort is improving choices of therapy and treatment for an even broader range of cancer types. This thesis focuses specifically on head and neck cancers, including the broad categories of nasal cavity and paranasal sinuses, larynx, pharynx, lip and oral cavity, salivary gland, and unknown primary head and neck cancers. However, the QIN includes many institutions that focus on different types and sites of cancer.

To date, efforts in using Machine Learning (ML) techniques in the medical field have been focused primarily on diagnosis. Early techniques included using 2-D imaging to detect features similar to some of the 3-D features that have been extracted by other efforts with PET-CT images. (Celebi, Kingravi, Aslandogan, & Stoecker, 2011) Some of the more recent efforts have reported using ML to predict survival in cancer patients including, but not limited to, a time to event prediction where a time to recurrence or death is predicted (Cruz & Wishhart, 2006), or a prediction of survival past a given time period (usually years) (Buchner, et al., 2013). The latter can be considered a multi-class ML problem, which requires a different set of algorithms and techniques.

Feature selection is utilized to obtain a set of features that is able to distinguish the outcome of the patients. Feature selection also assists in gaining an understanding of the relative informativity of factors that are important to outcome prediction. Understanding

the factors important to outcome prediction can help clinicians to focus time and energy on those factors. Feature selection is also helpful in assisting the evaluation of the scalability and effectiveness of ML methods for predicting outcome. Feature selection is an integral part of designing a system that is applicable to more than a single problem as each different disease, or different site of a cancer can affect predictive power. The process of selecting features is not as simple as choosing features that are believed by experts to be predictive, though domain knowledge may assist in evaluating the effectiveness of the selected features.

Outcome prediction is one step in attempting to provide a personalized treatment plan for each patient. When including treatment data and the PET scan immediately following the treatment, an outcome on both the effectiveness of the treatment as well as the final outcome in the patient's full cancer treatment may be obtained. When including treatment information, classifiers can potentially be used in a clinical decision support system. Clinical decision support systems are tools for clinicians to use when attempting to treat conditions that may have been seen before in the clinic, or for allowing clinicians to see similar patients to a current patient with previously unseen conditions.

The aim of this study is to discover whether the methods of outcome prediction are effective in accurately predicting outcome given only a partial set of the available data. The results presented in this thesis are based on a cohort of 147 patients with clinical information and images for at least two PET scans; one prior to treatment, and one after treatment. The set of patients studied contained 61 patients with cancer localized in the pharynx region. An accuracy of 80.33% (compared to a baseline majority classifier of 60.66%) was achieved utilizing this reduced cohort of 61 patients. Further, it is shown that this limited cohort has the power to provide valuable information on outcome prediction with just four to six features. Feature selection includes both clinical features and quantitative imaging features including the site of cancer, primary tumor volume, and race.

BACKGROUND

Machine Learning

The methods in this thesis research rely heavily on the use of machine learning (ML) techniques. There are two distinct modes for applying an ML approach to a set of data -- supervised (Kotsiantis, 2007) and unsupervised learning (Gentleman & Carey, 2008). Unsupervised learning, or clustering, is a ML technique where the goal is to assign labels (or classes) to previously unlabeled data. The unlabeled data is input to an algorithm where a data label or class is applied to each of the instances of a dataset. Because there is no known true value, unsupervised learning is useful when looking for trends in the data, however this final given clustering is not guaranteed to be a globally optimal solution, and may instead be just a single locally optimal solution.

Supervised ML is a technique that begins with labeled data where each datum is assigned to a unique class. Usually, the data is split into a training set and a testing set where the training set is used to construct rules or mathematical models that can then predict the class labels of the subset of data in the test class (with known, but masked labels). Supervised ML uses different mathematical models, called kernels, to build the classifiers which predict class labels for the members of the test dataset. The concept of a single training set and test set only illustrates the basic structure of the process used for constructing and validating classifiers. In practice, it is necessary to repeat the process of validation by subdividing the data numerous times. This repeated process is referred to as K-fold cross-validation. K-fold cross-validation involves partitioning the data into K sets where 1 of the sets becomes the test set and the remaining K-1 sets become the training set (Kohavi, 1995). When the value of K is equal to the number of instances (n) in the complete dataset, this method is called leave-one-out cross validation (LOOCV). Leave-one-out cross validation has the advantage of using the largest available training set in

order to attempt to classify a single test case, which gives the maximum number of instances from which to train. (Kotsiantis, 2007)

One common performance metric used for evaluating ML systems is area under the receiver operating characteristic curve (AUC) (Bradley, 1997). AUC provides a measure of the discriminatory power of a classifier for a given dataset. An AUC value of 0.5 means that the classifier performs no better than the flip of a coin. An AUC value greater than 0.5 indicates that a classifier has more discriminatory power than random guessing, while an AUC value less than 0.5 indicates a less-than random guessing performance of the classifier. In order to calculate an AUC value, both correct and incorrect guesses need to be made for each built classifier. Because of this property, no AUC metrics are available for LOOCV trials as each has either a correct guess or an incorrect guess, but not both.

In any ML problem, the choice of outcome needs to be made. In cancer, there are several different commonly used outcome categories or prognoses. (National Cancer Institute, 2012) The categories used in this study are disease-free survival, recurrence-free survival, and an optimal survival. Disease-free survival is a patient who is currently disease-free or has died from non-cancer related causes. Recurrence-free survival is a patient who falls into the disease-free survival outcome, but also has not had any recurrence of cancer. Optimal survival is designated as a patient who has been recurrence-free, and disease-free for more than 2 years. (Buatti, 2013) The two-year time point was chosen, since a recurrence event after the 2-year time point is more likely new disease and not a recurrence from the same cancer. For the purposes of this study, the optimal survival outcome was selected as the primary outcome class.

Machine Learning Algorithms

Many ML algorithms will be used, so background on each will be included to give a general understanding of their use. The ML algorithms used are logistic regression, radial basis function network, support vector machine, and random forest.

Logistic regression (Mitchell, 2010) is a ML algorithm that takes a vector of discrete (nominal) and continuous variables and gives a probability of the class given the vector of variables. Logistic regression is a generalization of linear regression, which is used primarily for predicting binary or multi-class dependent variables. Logistic regression generates a linear expression for classification, and the output is a probability. The choice of the value within the probabilities that classifies one class versus the other is calculated by the classifier. Logistic regression can function with data that is considered to be either conditionally independent or conditionally dependent. This property allows for many diverse datasets to be considered for classification with a logistic regression classifier. Even though it's not used in this study, logistic regression may also utilize a multi-class outcome variable.

The radial basis function network (RBF Network) is a subtype of artificial neural network that uses a linear combination of radial basis functions for interpolating the function which maps the variables or features to the class. (Broomhead & Lowe, 1998) Much like logistic regression, the RBF network outputs a numeric variable, which can determine a binary output by selecting a threshold value. The RBF network has the ability to group data by means other than a linear separator, for instance a cluster of data may be signified by a circular separator.

Support Vector Machines (SVM) (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998) employ a single hyperplane which attempts to separate the data. The hyperplane used to separate data can be calculated by one of many different kernels, a 2^{nd} order polynomial allows for separators to be of a non-linear and non-one-to-one scale. Between poly-2 kernels and RBF kernels, separation of data can be optimized as driven by the

data. Much like logistic regression, support vector machines also work with two or more classes for classification. Support vector machines work well in high-dimensional spaces, but if the number of features is much more than the number of samples, the performance of the classifier may be poor. The output of the support vector machine is not a probability like logistic regression was, but a score for each class or in the binary class case, a single score.

Random forest (Breiman, 2001) is a specific type of ensemble ML algorithm. In a random forest classifier, *N* decision tree classifiers are made with the intent that the data is run through all of these *N* classifiers and the final class for an example is based on one of a number of mechanisms. A few examples of the determination of the final class are a weighted average of the individual decision trees, or a voting majority of the individual decision trees. A random forest classifier is able to quickly and effectively process a dataset, and they work well with unbalanced or missing data, however random forests tend to over-fit the given training data.

PET Imaging

Positron Emission Tomography (PET) Imaging was used in conjunction with Computed Tomography (CT) imaging for diagnosis and progression of cancer for the patients. PET uses on an injected radioisotope, in the case of this study ^{18}C-fluorodeoxyglucose (FDG). FDG emits a positron that interacts with an electron in the body. That electron interaction emits two gamma rays that are detected by the PET detectors. The resulting image from a PET scan can be seen in Figure 1, for this patient, the tumor is located in the base of tongue. PET imaging is quantified by counting the number of coincident gamma ray hits in a specified amount of time (Society of Nuclear Medicine and Molecular Imaging). With the addition of CT, the hits that lie along a line between the two detectors that measured the coincident hit provide the ability to better calculate the location of the event. When run in conjunction with a CT scan, standardized

uptake values (SUV) may be calculated based on the gamma ray count, injected dose, and patient weight (Thie, 2004). The SUVs are used to calculate the tumor activity, and spatial metrics such as tumor volume. A common use of SUV is determining whether or not a tumor is benign or malignant. In general, an SUV of 2.5 is the cutoff between benign and malignant, but institutions may choose a different cutoff value. In the data used in this paper, the determination of benign or malignant was not made in favor of simply using the values obtained by the scan.

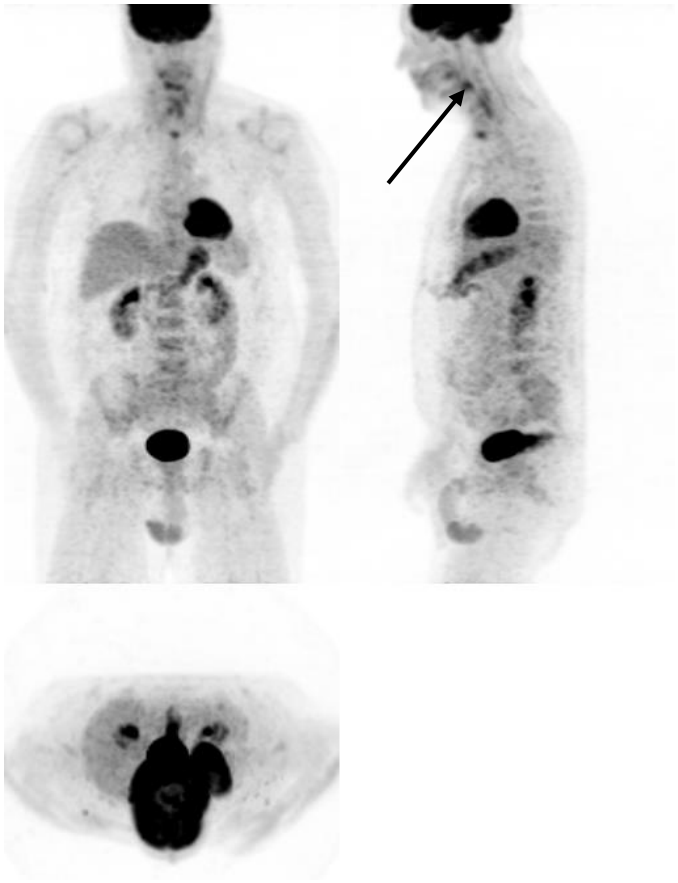

Figure 1: Example of the output of a PET-CT scan for one of the patients in the study, Arrow points to the location of the tumor.

METHODS

In this chapter the data, tools, and principle modes of analysis will be described, beginning with the source and repository of clinical data followed by the description of the goals and means of determining the predictive power of this data to contribute to improved clinical cancer care. A persistent contribution of this work is the establishment of a computational pipeline beginning with data extraction from an anonymized local research database, followed by feature selection, construction and tuning of a predictive classifier, and finally validation, and performance assessment.

Data Selection: Clinical Feature Pool

Clinical data was entered by clinicians into an anonymized local research database using a web front-end interface. Clinical information about each patient was entered directly from the clinical notes. The clinical features available to the classifier can be found in Table 1. Features that had categorical values list the possible values, while features with quantitative/continuous values give a range of the minimum to maximum value observable for the corresponding feature. The features in Table 1 are common to all datasets, and have identical values and ranges for all datasets that use these values. Features that were PET-specific in the clinical features (specifically PET hottest node maximum SUV, and PET maximum SUV) were the values provided by the clinicians reading (evaluating) the images at the time of the scan. PET status was indicated for each of the provided PET Scans, though only pre-treatment PET scans were used for the analyses reported here.

Table 1: Available clinical features

Feature	Range or Possible Values
Age at Diagnosis (years)	20.36 - 79.54
Chewing Tobacco Use	Yes, No, Former
Type of Diabetes	No, Insulin-Dependent, Non-Insulin-Dependent
Differentiation	Poor, Moderate, Well, Undifferentiated, In Situ, Not Available
Type of Drinker	No, Social, Significant
Site of Cancer	Tonsil, Oropharynx, Base of Tongue, Pyriform Sinus, Nasopharynx, Hypopharynx
Cancer State	2, 3, 4a, 4b
Node-Stage	0, 1, 2a, 2b, 2c, 3
Tumor-Stage	2, 3, 4, 4a, 4b
Gender	Male, Female
Height (cm)	135-200
Previous Radiation	Yes, No
Prior Malignancies	No, Prior Head & Neck, Prior Other
Race	Caucasian, Unknown, Asian, Native American
Smoker	Yes, No, Former
Weight (kg)	37-174
Body Mass Index	16.5-49.8
PET Hottest node SUV Max	0-25.4
PET Maximum SUV	0-37
PET Status	Abnormal, Normal, Equivocal

Dataset Selection: Quantitative Image Metric Pool

After clinical data was entered into the database, one of three tracers extracted the 3D PET-CT scans from the image database to perform tracings of the primary tumor for each of the patients in the dataset. Once the primary tumor region of interest (ROI) was selected and saved, analysis was performed on the PET scan to generate the quantitative indices in Table 2. In the PET modality, measures of uptake of glucose in tumors are referred to as standardized uptake values (SUV) and are a measure or count of radioactive decay from the radioactive tracers injected into the body normalized by weight. In addition to quantitative data from the primary tumor, automated localization and

segmentation of the aorta, cerebellum, and liver was performed on each PET scan to produce a mean SUV for those areas for normalization of the tumor-specific values.

Table 2: Table of the quantitative indices and ranges calculated for the primary tumor for each of the different tracers.

Feature	**graduate student Range**	**medical student Range**	**doctor Range**
Background Mean SUV	0.49 - 1.05	0.55-1.32	0.17 - 4.09
Mean SUV	3.84 - 11.97	3.65-10.67	3.19 - 11.53
Metabolic Tumor Volume	1946 - 701792	1946 - 711907	10036 - 696520
Peak SUV	4.10 - 29.18	4.10 - 29.18	4.07 - 29.127
Maximum SUV	4.19 - 29.44	4.19 - 29.44	4.19 - 29.44
Volume (mm^3)	506.9 - 77070.4	506.9 - 82554.6	2999 - 75811

The final dataset compiled for each of the three tracers includes the clinical features and the quantitative indices for the individual tracer. In total, each dataset has 61 instances and 30 features; 10 quantitative features from each image, 17 clinical features from the patient's first visit to the clinic and 3 additional clinical features for each PET scan. For the results presented here, only the pre-treatment PET scan was included.

Feature Selection

Using Weka, which has already implemented many ML techniques (Hall, et al., 2009), a logistic regression model was used to perform a sequential forward feature selection on each of the entire 61 patient datasets, as illustrated in Figure 1. Logistic regression was selected because it is better suited for multi-value nominal features, and its flexibility in assuming that features within the set of features can be either independent from, or dependent upon each other. (Mitchell, 2010) Experimental validation was performed using multiple classifiers, as each of the classifiers yield a different insight to the overall predictability of the set of selected features.

The flowchart in Figure 2 represents the protocol used for serial selection of best-performing features in compiling the set for experimental validation with the different ML kernels. This method is followed for each tracer's dataset. These datasets for tracers include the common clinical features, as well as the quantitative values from the tracings performed by the tracer. The features from each dataset were then separated with the desired outcome followed by 10 iterations of a 10-fold cross-validation classification by logistic regression. Ten iterations were performed to give a After the classifiers were built, the accuracy of the classifiers was used to either select the best performing feature, add it to the remainder of the features, or terminate the selection of additional features once the overall accuracy and performance ceased to improve.

Feature Set Validation

Feature sets were iteratively evaluated utilizing five different classifiers. A LOOCV analysis was conducted to provide results for comparison. For each dataset, both

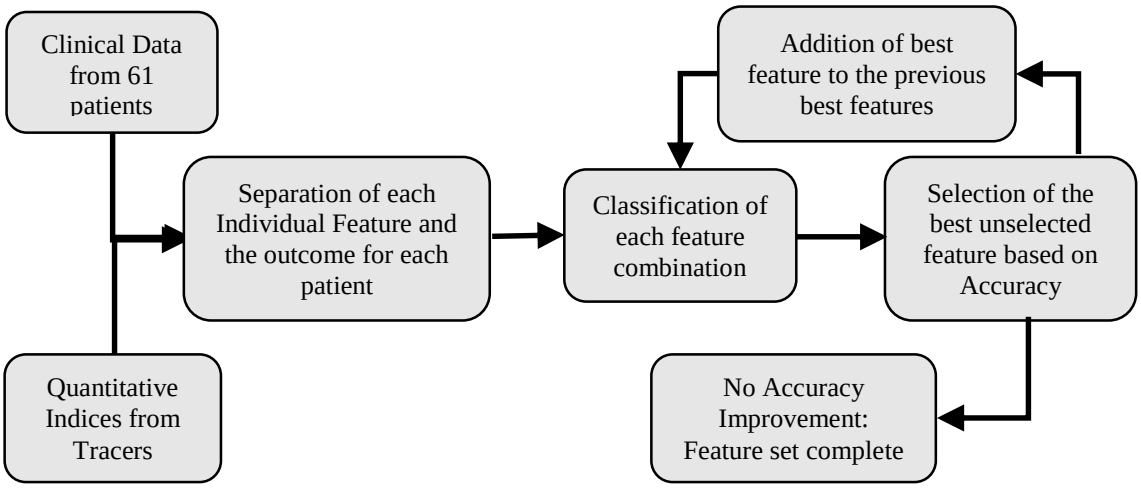

Figure 2: Flowchart of the sequential forward feature selection procedure where inputs included demographic, clinical, and quantitative features

the set of selected features and the set of all available features were processed by the

LOOCV analysis pipeline. Each set of features was processed by the same set of classifiers (including the majority classifier). Comparisons between the selected image-based features and all features were not performed, because each tracer used a different protocol to identify the image features. For instance, the graduate student and medical student selected regions of interest and a threshold between the background and tumor, while the doctor selected center points and had regions automatically filled in. However both the set of selected features from the sequential forward feature selection and the set of all available features were compared to the majority classifier.

Power Analysis

A power analysis was performed to evaluate whether increasing the sample size (of patients/cases) or decreasing intersample variance would increase the performance of the ML model. With data readily available, a post hoc power analysis was performed on all features for each tracer to determine future predictability with possible changes in the number of samples or intersample variance. For each of the features, four analyses were performed. Two analyses changed only the number of samples, keeping the interclass variation unchanged. The equation used for power is below.

$$z_{Power} = -1.96 + \frac{|\sigma_{yes} - \sigma_{no}|}{\sqrt{\frac{stdev_{yes}}{n_{yes}} + \frac{stdev_{no}}{n_{no}}}} \quad (1)$$

In the first analysis, the total number of samples was doubled, keeping the distribution of outcome unchanged. The second analysis increased the sample size, but created equal samples of optimal survival or not. The third analysis changed only the intersample variance of the values within the data while keeping the number of samples constant. For power analysis, this third change should give the same power value as doubling the sample size. The fourth analysis changed both the number of samples and interclass variation of the samples. Alternately, it may be that a change (lowering) in interclass variation is sufficient, or that a combination of both is needed to obtain higher

accuracies in classification. A 95% confidence interval was selected for analysis to compute power using a significance of $p < 0.05$, this is the -1.96 value in the equation above. (Thomas, 1997)

Simulated Analysis

In an effort to get a preliminary result from the power analysis, data was simulated in a manner to match the potential change in both number of samples and intersample variance. The simulation was taking the data that was already captured and used and either copying the data for more instances or changing values of data in a equal manor to reduce the deviation for those values. Based on the power analysis performed previously, all three tracer's data was combined in a manner to create a dataset that provides the closest dataset to the dataset that was modeled by the power analysis. After instances in all three datasets with no quantitative imaging indices were removed (three from the yes class, two from the no class) a weighted average of all three tracers was calculated. The weighting was 4:2:1 for the doctor, medical student, and graduate student respectively. This meant that the doctor was weighted twice as much as the medical student, who was weighted twice as much as the graduate student based on expertise and experience in the head and neck region. After the weighted average dataset was calculated, Figure 3, then processed through an operation of reducing the standard deviation by dividing the difference between the value for the example and the mean for its class by two. This last operation did not change the mean value per class, only the interclass standard deviation was modified. After the dataset was built, it was put through the same sequential forward feature selection as the original datasets, it was then run through the experimental validation. An additional dataset was created by duplicating all of the examples effectively doubling the number of examples in the dataset.

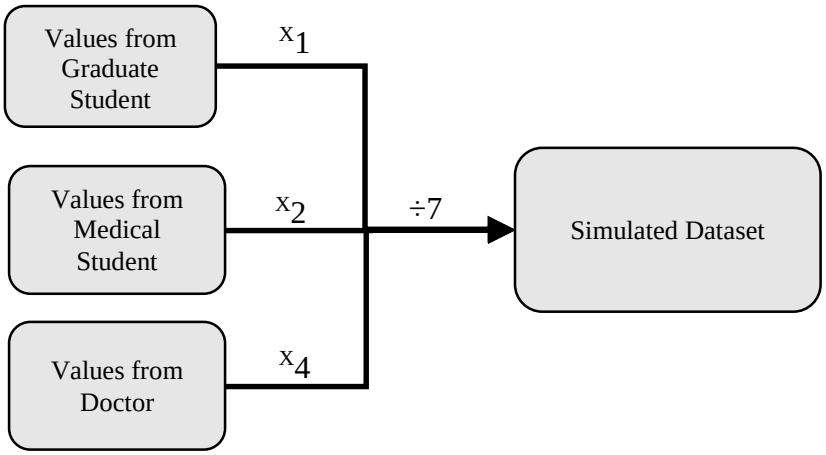

Figure 3: Flowchart of the makeup of the simulated dataset, which normalizes the data from all 3 tracers' datasets

RESULTS

Each of the tracer datasets were processed by an identical pipeline to produce the list of the top performing features as outlined in the previous section. Features were selected based on their relative performance when combined with previously selected features using a 10-run 10-fold cross-validation logistic regression model. If there was a tie between two features, the feature with a higher AUC (Area under the ROC Curve) was selected. Between the three different datasets when trying to predict the optimal survival outcome, commonly selected features included tumor site and whether or not the patient had diabetes (and which type they had). Tumor volume was a selected feature for both the medical student and the graduate student, but not the doctor. While the accuracy during feature selection was important to the process of selecting features, the values obtained would be invalid for using as a final measure of the performance for a given set of features. Performance in feature selection is invalid for measuring overall performance for a given feature set because each iteration of the selection process uses the same data subset partitions. Therefore, there is a tendency to over-fit to those specific partitions.

Feature Selection

Table 3 provides an overview of the features that were selected, the overall accuracy and for which tracer the feature was selected for. The number indicates the order of selected features for the given tracer. The highest accuracy indicates the accuracy achieved with all selected features for the tracer. Table 3 also shows which features are indicative across all tracers and what order that feature was selected for all tracers. Cancer site and diabetes were selected by all 3 tracers where drinker, background SUC and tumor mean SUV were selected by only one tracer.

Table 3: Features with feature at which the given feature was selected

Feature	graduate student	medical student	doctor
Cancer Site	1	1	1
Race	2		2
Height	3		3
Diabetes	4	3	4
Tumor Volume	5	2	
Weight	7	4	
Drinker		5	
Background SUV		6	
Tumor Mean SUV	6		
Highest Accuracy	80.48%±15.18	81.57%±16.13	78.74%±14.73

Figure 4 shows the percent change in accuracy from the majority class based on the number of features selected. The selected features are listed above in Table 5 with the value of the highest accuracy being the last point for each of the tracer's graphs. The doctor and the graduate student match exactly for the first four points as the first four features selected by both are the same and have the same order. The graduate student then goes on to select more features, the doctor does not. Since all three tracers had a different number of selected features, the lines stop for each tracer after all selected features were completed. Since all first selected features were the same, the first point was the same for all 3 tracers.

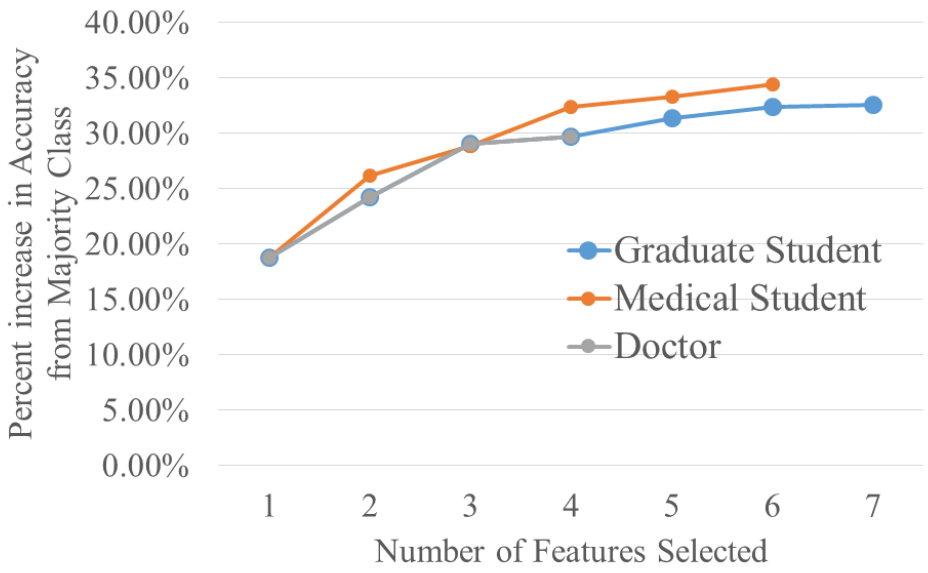

Figure 4: Plot of percent increase in accuracy vs the number of features selected for the tracers

Feature Set Validation

In addition to utilizing a 10-run, 10-fold cross-validation for feature selection, a leave-one-out cross validation (LOOCV) was performed in order to test the validity of the feature selection that was performed using multiple classifiers. While logistic regression shows significant improvement with the set of selected features, other classifiers had similar significant increases in accuracy with the same sets of features. There were two major comparisons to be made for this dataset; the first comparison was with only the selected features against the majority classifier, and the second was with no feature selection against the majority classifier. With these two comparisons, the effectiveness of the feature selection could be tested. Both of these comparisons are presented in Table 4.

Table 4: Classifier performance for each tracer

Classifier/Features	**graduate student**	**medical student**	**doctor**
Support Vector Machine (LOOCV)			
Majority Classifier	60.66%	60.66%	60.66%
No Feature Selection	59.02%	59.02%	59.02%
Selected Features	73.77% *	68.85% *	72.13% *
Logistic Regression (LOOCV)			
Majority Classifier	60.66%	60.66%	60.66%
No Feature Selection	42.62%	44.26%	42.62%
Selected Features	77.05% *	80.33% *	78.69% *
RBF Network (LOOCV)			
Majority Classifier	60.66%	60.66%	60.66%
No Feature Selection	64.43%	63.61%	61.64%
Selected Features	73.11%	70.49%	75.25% *
Random Forest (25) (LOOCV)			
Majority Classifier	60.66%	60.66%	60.66%
No Feature Selection	57.38%	61.48%	58.36%
Selected Features	74.26%	73.44%	72.30%

NOTE: in table * indicates $p<0.05$ when compared to Majority Classifier

The data in Table 4 show that the doctor's annotations, which only had features common to all datasets as the selected features, was the only dataset to have significance from the majority classifier with the RBF network. The random forest classifier yielded no significant increase over the majority classifier.

Table 5: List of features selected and the number of occurrences of that feature being selected out of the three possible datasets.

Feature	Number of datasets where feature was selected
Site of Cancer	3
Diabetes	3
Race	2
Height	2
Tumor Volume	2
Weight	2
Drinker	1
Background SUV	1
Tumor Mean SUV	1

Table 5 provides a list of all selected features and the number of datasets in which that feature was selected by the feature selection process. Notably, tumor volume which is a dataset-specific feature, was selected by more than one dataset. Site of cancer, which

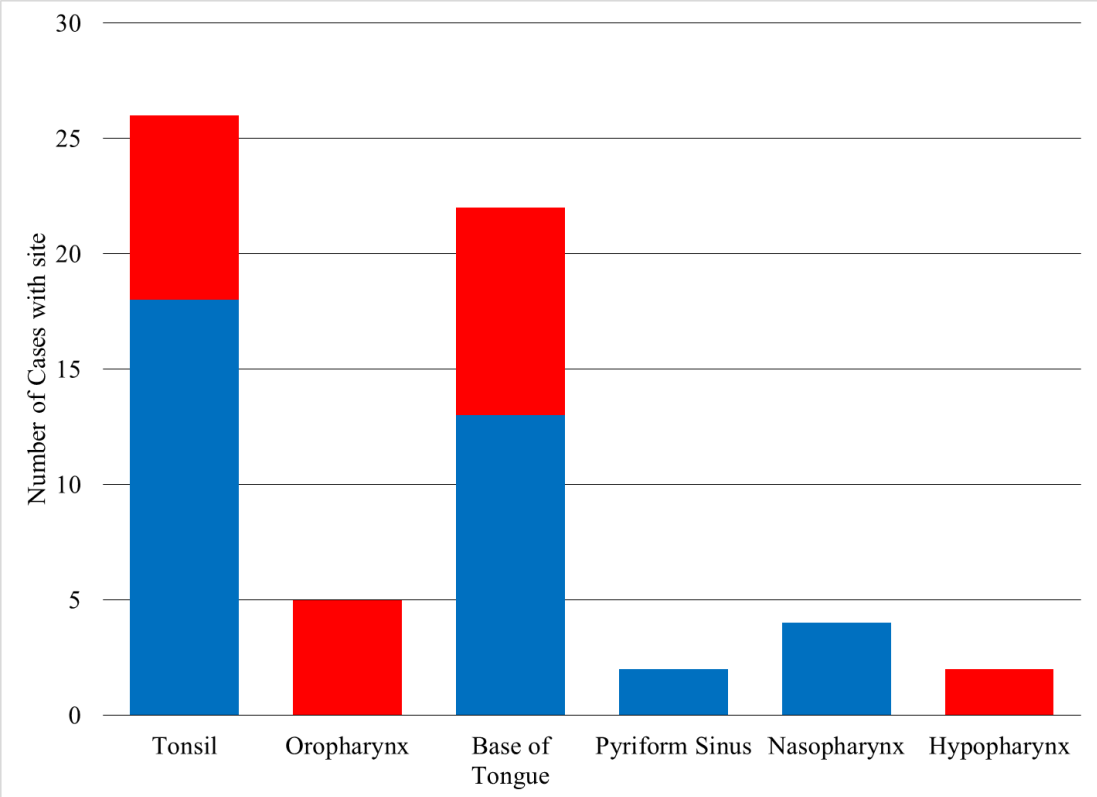

Figure 5: Breakdown of outcome by site, good outcome (2-years cancer-free with no recurrence) in blue, and bad outcome in red.

was selected in all three datasets, had many possible values, and can be seen in Figure 5. Also seen in Figure 5 is that the four rarest sites have only one outcome class.

Power Analysis

A power analysis as described earlier was performed on each tracer's available quantitative features. This analysis used a 95% confidence interval, post hoc, using the current data from the tracers as the source for mean and standard deviation values. The objective of this power analysis was to determine the potential usefulness of enhanced data (a larger total dataset size or an improved S/N ratio) in improving classifier performance. In Table 6, the steps are shown which were used to produce different values of power by artificially changing parameters. When either increasing the number of instances and decreasing the intersample variance in the samples, there is an increase in power. When both changes occur together, the power increases more than with either factor considered alone. Indeed this is a reasonable expectation for features such as tumor volume, as it is likely that both effects will occur together.

Table 6: Power Analysis Example with Tumor Volume

Change from observed results	**graduate student**	**medical student**	**doctor**
No Change	45.2%	40.6%	50.2%
Double Instances / Half Variation	73.9%	68.3%	79.4%
Normalize Instances	67.2%	61.9%	73.0%
Half Variation Double Instances	95.7%	93.1%	97.6%

By doubling the number of instances per feature, or doubling the size of the study, two of the features achieved a power greater than 80%. These two features were weight and body mass index, which are common to all three datasets. Further analyses included normalization of the number of samples for each of the possible outcomes. In the unbalanced dataset, the majority classifier contains 37 instances while the minority class

contains 24 instances. This power analysis normalizes the number of instances for both outcome classes to be 47 instances, rather than the 37 or 24 instances for the majority and minatory class respectively, while keep the intersample variance constant.

Table 7: Change in Power for the doctor's features

Feature	Empirical Power	Power with ½ intersample variance and 2x instances
Weight	70.76%	99.89%
Height	28.99%	80.30%
Body Mass Index (BMI)	53.06%	98.27%
Aorta Mean SUV	4.77%	8.46%
Cerebellum Mean SUV	3.93%	5.96%
Liver Mean SUV	7.75%	18.82%
Peak SUV	4.83%	8.63%
Maximum SUV	4.57%	7.84%
Metabolic Tumor Volume	29.99%	81.88%
Tumor Volume	50.24%	97.57%

Table 7 provides an example of some of the selected features and their power values for the doctor's tracings of the images, in the second column is the power of the feature at the time of analysis. In the final column is the power of the feature with the standard deviation reduced by one-half and the number of instances doubled. The first three features (weight, height and BMI) are common to all three datasets, while the last four features are specific to each individual tracer. The middle three features, aorta, cerebellum and liver mean standardized uptake value (SUV) are Peak SUV, maximum SUV, metabolic tumor volume, and tumor volume are all metrics that are indicated by clinicians and clinical notes. The medical student, on top of the features common to all tracer, had an increase in power for metabolic tumor volume (99.20%), tumor volume (99.82%), and mean SUV (87.97%) indicated to provide power to analysis. The graduate student had metabolic tumor volume (97.9%), background mean SUV (93.81%), and

tumor volume (99.94%) increase to a high power on top of the common features, weight, height, and BMI.

Simulated Analysis

Table 8 provides the results of the sequential forward selection for the dataset that contains the duplicated examples from the dataset that contains the reduced interclass standard deviation after weighted averaging. With an overall accuracy of 90.07% from a Majority Classifier of 58.94%, this is the largest increase in accuracy of any dataset after feature selection. While some of this accuracy can be attributed to simply duplicating the examples and the classifiers possibly having a duplicate example.

Table 8: Feature Selection with duplicate examples and half interclass standard deviation

Feature	**Accuracy**	**AUC**
Majority Classifier	58.94% ± 4.10%	0.50 ± 0.00
Tumor Volume	72.08% ± 10.16%	0.78 ± 0.15
Cancer Site	82.02% ± 9.74%	0.88 ± 0.10
Node Stage	84.64% ± 10.32%	0.90 ± 0.09
Metabolic Tumor Volume	90.07% ± 8.44%	0.95 ± 0.08

Table 9 provides the selected features using the sequential forward feature selection using the reduced interclass standard deviation of the dataset with weighted averages. This dataset contains no duplicate examples and still provides a higher accuracy than any of the individual datasets. The AUC is also about on par with the AUC of the medical student, which was the highest performing of the three tracers.

Table 9: Feature Selection with half interclass standard deviation and no duplicate examples

Feature	**Accuracy**	**AUC**
Majority Classifier	58.94% ± 4.10%	0.50 ± 0.00
Tumor Volume	71.13% ± 16.01%	0.75 ± 0.21
Cancer Site	81.40% ± 14.85%	0.89 ± 0.15
Drinker	82.93% ± 15.27%	0.87 ± 0.19
Tumor Mean SUV	84.13% ± 15.53%	0.86 ± 0.20
Previous Radiation	84.30% ± 15.43%	0.87 ± 0.20

Table 10 provides the results from validation of the selected features compared to all features for both datasets from the simulated analysis data. The dataset containing duplications will be guaranteed to have the instance being tested in the training data with LOOCV, so the results are skewed for the dataset containing duplicates. For the dataset without duplicates, such as Table 3, an observation can be made that the feature selection gives a higher performance than the dataset with no feature selection. In the simulated datasets, logistic regression still holds the highest accuracy, and only Support Vector Machine and logistic regression have significant increases from the majority classifier.

Table 10: Classifier performance for the two created simulated datasets

Classifier/Features	**Accuracy without Duplicates**	**Accuracy with Duplicates**
Support Vector Machine (LOOCV)		
Majority Classifier	58.93%	58.93%
No Feature Selection	58.93%	81.96% *
Selected Features	69.64% *	78.57% *
Logistic Regression (LOOCV)		
Majority Classifier	58.93%	58.93%
No Feature Selection	66.07%	100.00% *
Selected Features	83.93% *	91.07% *
RBF Network (LOOCV)		
Majority Classifier	58.93%	58.93%
No Feature Selection	62.32%	99.91% *
Selected Features	72.86%	98.93% *
Random Forest (25) (LOOCV)		
Majority Classifier	58.93%	58.93%
No Feature Selection	61.79%	92.86% *
Selected Features	69.64%	82.14% *

NOTE: in table * indicates $p<0.05$ when compared to Majority Classifier

DISCUSSION

This study suggests that feature selection is an important step that will be helpful for this and other datasets in the future. The features selected by the doctor contain no dataset-specific quantitative imaging information. While all three tracers used the same program to perform the tracings, each tracer used a slightly different protocol when doing so. The doctor and graduate student were both able to use clinical notes to guide and verify the location of the primary tumor. Both the graduate student and the medical student selected the tumor by selecting a point in the tumor and expanding a region to cover the entire tumor and then manually choosing the threshold between the background and the tumor on the PET scan. The doctor selected points representing the center of the tumor. The region of interest was automatically selected by an algorithm that was integrated into the software used for selection. All tracers were required to select a background region of similar tissue for later normalization. All regions of interest were then segmented and indices calculated in the same manner for all three tracers and uploaded to the local clinical research database.

As each tracer used a slightly different protocol to select the primary tumors, the comparisons between the different tracers are challenging. The data used in the study was a subset of a larger set of data with more cancer sites. In this dataset, the site is limited to the pharyngeal region including tonsil, base of tongue, pyriform sinus, nasopharynx and hypopharynx. The tumor staging was also limited to stages 2, 3, 4a, and 4b as these sites and stages have similar treatment protocols. While the treatment and post-treatment information was not used for predicting an outcome, the selectivity of the data used assisted in limiting some of the variables that were not used for prediction. This selectivity of the data also results in a bias in data, as there are no examples of patients with a stage lower than 2, nor is there an instance of cancer site in a region other than the pharynx. Looking at Table 2, a distinction between the minimum and maximum of the

data corresponding to the 3 different methods can be made. This is most notably seen in the volume calculations, metabolic tumor volume and tumor volume, where the doctor's minimum values are an order of magnitude greater than the volume values of either the graduate student or medical student.

The results of the feature selection show the potential promise of imaging features that may be helpful in predicting outcome based on clinical and quantitative imaging data. For two of the datasets selected, the volume of the tumor was shown to be a predictive feature, which is indicative of the tumor size being a predictive feature of the outcome for patients with head and neck cancer. Other selected imaging features (from Table 5) include the background mean standardized uptake value (SUV), and the primary tumor mean SUV. The background SUV has the potential to be a term for which all mean SUVs are normalized against along with the aorta, cerebellum and liver mean SUVs that were automatically located, segmented and analyzed by C. Bauer et al. (Bauer, et al., 2012). While many features selected may appear to be obvious, some features, such as weight, height, diabetes, and drinker may appear counter-intuitive to experienced oncologists. As with the site of cancer, seen in Figure 5, many of the listed features have a tendency to have values with a small number of examples – each with the same outcome. Indeed, they look similar to Hypopharynx as seen in Figure 5, where it is the minority class and it can add a small number of properly classified cases. The quality of the completed tracings may be a factor affecting the accuracy of the findings, since the graduate student and medical student do not have as much domain knowledge as the radiation oncologist (doctor) and the radiation oncologist did not perform the full tracing procedure. The major question with the graduate student and medical student are whether the imaging indices are appropriately labeled. No validation of tracing location has been performed, and for this reason no cross-dataset comparisons were made.

The experimental validation adds credence to using the SVM classifier with the logistic regression feature selection. However, the random forest classifier had no

significant improvement over a simple majority classifier. With the selected features, performance was similar to the performance on the RBF network, as seen in Table 4, for the graduate student and medical student. Logistic regression experimental validation showed significant accuracy improvements over the majority classifier when using a set of selected features. Of the four algorithms evaluated, two (SVM and logistic regression) demonstrated significant improvements after performing feature selection.

One of the pitfalls of working with clinical data is the potential for inconsistent data quality. Much time has been spent in attempting to validate the data that was manually entered into the local research database. With a pre-existing model for the determination of cancer prognosis with the staging system (when stage was not present in the list of selected features) questions as to the validity of the feature selection process arose. There was a data bias intrinsic to the dataset, where outcome segregation of cancer stages was found to be similar to the outcome segregation of tonsil and base of tongue in Figure 5. While this bias does not reflect what is seen throughout the whole clinic, it represents the population of subjects that would be analyzed in this manner to obtain quantitative imaging indices. In the study, there is a bias towards patients that have higher stage cancers, as lower stage cancers are generally not imaged. The dataset also contained less instances of patients who had more severe cancers, as the instances with the higher cancer severity rarely had post-treatment PET scans. As a result, the instances that were included in the dataset had a better outcome percentage than was clinically observed because they had a post-treatment PET scan.

The power analysis provides some insight into the future direction of the predictability of the dataset with more data, and with data that is more coherent and consistent. With the already reduced set of patients from the original sample of patients, it is likely that more patients will be added later. Many features that were analyzed for power had a large increase in power by doubling the number of instances for both classes or by halving the interclass variance. The features common to all three tracers (weight,

height, and BMI), either double in power or obtain a power value close to 100%. For the common features, doubling the number of instances provided a doubling effect for the power. With a decrease in inter-sample variance, there was more power observed in the quantitative imaging features and even more power with an increase in the number of instances along with the decrease in inter-sample variance. The feature selection and power analysis showed that tumor volume was both important in predicting the optimal outcome in the study, and a very powerful feature for separating the two classes. Weight was another feature that was selected during feature selection and very powerful as indicated by the power analysis and was the most powerful of all of the features before modification of instance number or interclass variance with a power of 70.76%. For the doctor, the least powerful quantitative imaging feature was the tumor mean SUV, though the power doubled with changes in the number of instances and interclass variance. All features showed an increase in power when increasing the number of samples, which indicated that more samples would assist in increasing the future significance of the features. With a higher significance in the features based on power, the predictability of outcome with classifiers will likely be increased as the classifiers will be able to advantage of the increased separation of the outcome classes.

The simulated data analysis provided some validation to the power analysis. While the dataset that contained duplicate instances which were a leak of the final outcome, especially with LOOCV, the reduction in interclass standard deviation was sufficient to increase the accuracy beyond values obtained by the individual tracers. Similar features were selected in the simulated dataset to what was selected in the individual tracer's dataset. Tumor volume and cancer site were both selected in both the simulated dataset and in most of the tracer's individual datasets. The changes made in the simulated datasets only affected the quantitative imaging features specific to the tracers, which accounts for only 6 of the 29 total features. Though the changes only affected 6 of the features, some of the unaffected features continued to provide an increase in accuracy

that was more than had previously been achieved in the individual tracer's datasets. The accuracy of the dataset containing duplicates was indicative of the possibilities of the classifier with the sequential forward selection results presented in Table 8, while the dataset containing no duplicates was indicative of what some increased attention to the data, most specifically the image tracings, could do to assist in current outcome prediction.

REFERENCES

American Cancer Society. (2013). Cancer Facts & Figures 2013. Atlanta: American Cancer Society.

Bauer, C., Sun, S., Sun, W., Otis, J., Wallace, A., Smith, B. J., . . . Beichel, R. (2012). Automated Measurement of Uptake in Cerebellum, Liver, and Aortic Arch in Full-body FDG PET/CT Scans. Medical Physics, 3112-3123.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1145-1159.

Breiman, L. (2001). Random Forests. Berkeley: University of California Berkeley.

Broomhead, D., & Lowe, D. (1998). Multivariable functional interpolation and adaptive networks. Complex Systems, 321-355.

Buatti, J. (2013, October 28). Cancer Outcome Conversation. (D. Dellsperger, Interviewer)

Buchner, A., May, M., Burger, M., Bolenz, C., Herrmann, E., Fritsche, H. M., . . . Bastian, P. J. (2013). Prediction of outcome in patients with urothelial carcinoma of the bladder following radical cystectomy using artificial neural networks. The Journal of Cancer Surgery, 372-379.

Celebi, M. E., Kingravi, H. A., Aslandogan, Y. A., & Stoecker, W. V. (2011). Detection of blue-white veil areas in dermoscopy images using machine learning techinques. Signal Processing and Information Technology, 1960291.

Cruz, J. A., & Wishhart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 59-77.

Gentleman, R., & Carey, V. J. (2008). Unsupervised Machine Learning. In F. Hahne, W. Huber, R. Gentleman, & S. Falcon, Bioconductor Case Studies (pp. 137-157). New York: Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1).

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support Vector Machines. Intelligent Systems and their Applications, IEEE, 18-28.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In I. G. Maglogiannis, Emerging Artificial Intelligence Applications in Computer Engineering (pp. 3-24). Amsterdam: IOS Press.

Mitchell, T. M. (2010). Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. In T. M. Mitchell, Machine Learning (pp. 1-17). Boston: McGraw Hill.

National Cancer Institute. (2012, May 11). Understanding Cancer Prognosis. Retrieved from National Cancer Institute: http://www.cancer.gov/cancertopics/factsheet/Support/prognosis-stats

National Institute of Health. (2013, August). Quantitative Imaging for Evaluation of Responses to Cancer Therapies. Retrieved from National Cancer Institute: http://imaging.cancer.gov/programsandresources/specializedinitiatives/qin

Society of Nuclear Medicine and Molecular Imaging. (n.d.). PET Scans: Get the Facts. Retrieved from Society of Nuclear Medicine and Molecular Imaging: http://www.snm.org/index.cfm?PageID=7988

Thie, J. A. (2004). Understanding the Standardized Uptake Value, Its Methods, and Implications for Usage. THe Journal of Nuclear Medicine, 1431-1434.

Thomas, L. (1997). Retrospective power analysis. Conservation Biology, 276-280.