Response Assessment and Prediction in Esophageal Cancer Patients via F-18 FDG

PET/CT Scans

by

Kyle J. Higgins

Graduate Program in Medical Physics
Duke University

Date:_____

Approved:

_____
Q. Jackie Wu, Supervisor

_____
Shiva Das

_____
Brian Czito

_____
Bastiaan Driehuys

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Graduate Program in
Medical Physics in the Graduate School
of Duke University
2015

ABSTRACT

Response Assessment and Prediction in Esophageal Cancer Patients via F-18 FDG PET/CT Scans

by

Kyle J. Higgins

Graduate Program in Medical Physics
Duke University

Date:_____

Approved:

_____
Q. Jackie Wu, Supervisor

_____
Shiva Das

_____
Brian Czito

_____
Bastiaan Driehuys

An abstract of a thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Graduate Program in
Medical Physics in the Graduate School
of Duke University
2015

Abstract

Purpose: The purpose of this study is to utilize F-18 FDG PET/CT scans to determine an indicator for the response of esophageal cancer patients during radiation therapy. There is a need for such an indicator since local failures are quite common in esophageal cancer patients despite modern treatment techniques. If an indicator is found, a patient's treatment strategy may be altered to possibly improve the outcome. This is investigated with various standard uptake volume (SUV) metrics along with image texture features. The metrics and features showing the most promise and indicating response are used in logistic regression analysis to find an equation for the prediction of response.

Materials and Methods: 28 patients underwent F-18 FDG PET/CT scans prior to the start of radiation therapy (RT). A second PET/CT scan was administered following the delivery of ~32 Gray (Gy) of dose. A physician contoured gross tumor volume (GTV) was used to delineate a PET based GTV (GTV-pre-PET) based on a threshold of >40% and >20% of the maximum SUV value in the GTV. Deformable registration was used in VelocityAI software to register the pre-treatment and intra-treatment CT scans so that the GTV-pre-PET contours could be transferred from the pre to intra scans (GTV-intra-PET). The fractional decrease in the maximum, mean, volume to the highest intensity 10%-90%, and combination SUV metrics of the significant previous SUV metrics were

compared to post-treatment pathologic response for an indication of response. Next for the >40% threshold, texture features based on a neighborhood gray-tone dimension matrix (NGTDM) were analyzed. The fractional decrease in coarseness, contrast, busyness, complexity, and texture strength were compared to the pathologic response of the patients. From these previous two types of analysis, SUV and texture features, the two most significant results were used in logistic regression analysis to find an equation to predict the probability of a non-responder. These probability values were then used to compare against the pathological response to test for indication of response.

Results: 20 of the 28 patients underwent post treatment surgery and their pathologic response was determined. 9 of the patients were classified as being responders (treatment effect grade ≤ 1) while 11 of the patients were classified as being non-responders (treatment effect grade > 1). The fractional difference in the different SUV metrics has shown that the most commonly used maximum SUV and mean SUV were not significant in determining response to the treatment. Other SUV metrics however did show promise as being indicators. For the >40% threshold SUV to the highest 10%, 20%, and 30% (SUV10%, SUV20%, SUV30%) were found to significantly distinguish between responders and non-responders ($p=0.004$) and had an area under the Receiver Operating Characteristic curve (AUC) of 0.7778. Combining these significant metrics (SUV10% with SUV20% and SUV 20% with SUV30%) also was able to distinguish response ($p=0.033$, AUC=0.7879). Cross validation of these results shown that these

metrics could be used to find the response on previously unseen data. The three individual SUV terms distinguished responders from non-responders with a sensitivity of 0.7143 and a specificity of 0.6400 from the cross validation. Cross validation yielded a sensitivity of 0.8333 and a specificity of 0.7727 for the combination of SUV10% and SUV20% and a sensitivity of 0.8333 and specificity of 0.7273 for the combination of SUV20% and SUV30%. For the >20% threshold two SUV metrics were found to be significant. These were the SUV to the highest 10% and 20% (p=0.0048). The AUC for the 10% metrics was 0.7677 and for the 20% metric it was 0.7374. Cross validation of these two metrics shown that the 10% metric was the better indicator with being able to distinguish response in unseen data with a sensitivity of 0.7778 and a specificity of 0.7727.

The only texture feature that was able to determine response was complexity (p-0.04, AUC=0.7778). This metric was no more significant than the three individual SUV metrics but less significant than both of the combination metrics. As with the SUV metrics, cross validation was able to show the robustness of these results. Cross validation yielded a result that could accurately distinguish a response with a sensitivity of 0.8333 and a specificity of 0.7273. Logistic regression fit with features of the two most significant results (complexity and combination of SUV10% with SUV20%) yielded the most significant result (p=0.004. AUC=0.8889). Cross validation of this model resulted in

a sensitivity of 0.7982 and a specificity 0.7940. This shows that the model would accurately predict the response to unseen data.

Conclusions: This study revealed that previously used SUV metrics, maximum and mean SUV, may have to be rethought about being used to determine a response in esophageal cancer patients. The most promising SUV metric was a combination of the SUV10% and SUV20% metric for a GTV created from a threshold of >40% of the maximum SUV value, while the most significant texture feature was complexity. The overall best indicator was the logistic regression fit of the significant metrics of complexity and combination of SUV10% with SUV20%. This was able to distinguish responders from non-responders with a threshold of 0.3186 (sensitivity=0.9091, specificity=0.7778).

Contents

List of Tables

List of Figures

Acknowledgements

Shiva Das, PhD

Q. Jackie Wu, PhD

Manisha Palta, MD

Brian Czito, MD

Bradford Perez, MD

Christopher Willett, MD

Jeff Nawrocki

1. Introduction

Esophageal cancer is a serious and aggressive malignancy and typically it appears in two forms, squamous cell carcinoma and adenocarcinoma. Squamous cell carcinoma (SCC) is the predominant type of cancer that appears in the esophagus worldwide. The incidence of SCC tends to increase with age and is commonly found more often in people of color, while adenocarcinoma is more prominent in Caucasians [1]. The aggressive nature of the disease leads it to being one of the top cancers causing mortality in patients.

1.1 Occurrence of Esophageal Cancer

According to recent SEER reports, in 2014 there was an estimated 18,170 new cases of esophageal cancer occurred in the United States. This accounted for roughly 1.4% of all new cases of cancer over all. Along with these new cases, the disease claimed an estimated 15,450 lives, accounting for roughly 2.6% of all cancer deaths that year. As can be seen, esophageal cancer is not very common in the United States. In fact, it ranks eighteenth among newly occurring cancers in the U.S. It also ranks as the tenth leading cause of cancer death [2].

Globally it ranks as the eighth most commonly occurring type of cancer with an estimated 462,000 new cases being recorded in 2002 which was 4.2% of the total new cases overall. It was also the sixth leader in cancer related deaths causing 386,000 fatalities, which was 5.2% of the overall deaths caused by cancer [3]. Overall esophageal

cancer is more common and more deadly outside of the United States. This could be the result of less effective healthcare in developing countries which do not have the money or resources for the best treatment opportunities.

1.2 Survival Rate of Esophageal Cancer

While the disease tends to be among the lower percentage of occurrences and deaths caused by cancer, the survival of patients who are diagnosed is low. In the United States the overall 5-year survival rate is 17.5%. The survival of the patient largely depends on the stage and localization of the disease. If the tumor is confined to where it has originated the 5-year survival rate of the patient increases to 39.6%. If the disease has regional metastasis, the prognosis decreases to a 21.1% 5-year survival rate. The worst case for the patient would be if the disease had distant metastasis, which would result in an overall 5-year survival rate of just 3.8% [2]. The difference in the stage of the disease could determine what type of treatment options there are for a patient.

1.3 Local Tumor Failure with Current Treatment Techniques

Current treatment strategy for esophageal cancer is a combination of chemotherapy and radiation therapy followed by surgery. This has led to an increase in the survival of patients suffering the disease. However, many patients will forgo surgery by either declining to have it or not being able to physically tolerate the procedure. For these patients that do not receive the surgery, chemoradiation is the standard approach. Despite improvements in the delivery in radiation therapy with intensity modulated

radiation therapy (IMRT) and image guided radiation therapy (IGRT), local tumor failure still occurs in patients that suffer from esophageal cancer.

A study done by Welsh et al looked at the patterns of these local tumor failures. The study monitored 239 patients, of which 119 (50%) had local reoccurrence of the disease and 116 (48%) had distant tumor failure outside of the radiation field as well. Most of the patients that had the local tumor failure (107 of the 119), did so in the gross tumor volume (GTV) designated by a physician. Failure also occurred in the clinical target volume (CTV) and in the planning target volume (PTV). Eighty-six patients only had local tumor failure in the GTV [4]. The GTV is an anatomical structure that is drawn on diagnostic images by a physician where he believes the actual size and shape of the tumor is. The study shows that the majority of the patients had failures in this region after the completion of the treatment.

1.4 Need for Measuring Response of Treatment

With local tumor failures being prominent in patients that suffer from esophageal cancer, there is a need for a measurement and prediction of response of the patient to the treatment. With a measurement of response mid-treatment, a prediction could be made of the outcome of whether or not there will be a tumor failure. With this prediction, the treatment could then be reassessed and altered to better benefit the patient's well-being.

1.4.1 Current Commonly Used Measurements of Response

Currently, the most common technique for measuring the treatment response of patients is by utilizing PET/CT images. A PET/CT scanner is an imaging device that has both a CT scanner and a PET scanner mounted on the same machine. This is beneficial since it will limit patient movement between the scans allowing for accurate comparisons. When administering a PET scan, a radiotracer is used as a contrast agent on the image. Currently for the measurement of response to esophageal cancer the most commonly used radiotracer is F-18 flourodeoxyglucose (FDG). This is a sugar analog that enters the body and is easily taken up into the blood and delivered to tumors since they tend to be very vascular. A measurement of how well a region of the body takes up this radiotracer is done so by the standard uptake value (SUV) which is described more in later sections. This SUV of the tumor is what is currently being used to indicate response.

The most commonly used SUV parameters in current studies are the change in the SUV max and the SUV mean [5] [6] [7]. These values are found by analyzing the F-18 FDG PET/CT image's voxels in the region of the disease site for scans that are taken both prior to the treatment and then about midway through the treatment. In the case of the maximum SUV metric, the voxel in the disease region that has the highest SUV is recorded and then compared to the highest SUV voxel in the intra-treatment scan. The mean SUV is found by averaging the SUV voxels in the region of the disease.

Although both the maximum SUV and the mean SUV have been cited in a multitude of studies, there has been no clear indication for which of these SUV parameters would be the overall best indication for response. Furthermore, there has been a lack of studies that test SUV parameters outside of the regularly used maximum and mean SUV metrics. There is a need for further study as to what metric could be used for a predictor of the response to the treatment of esophageal cancer patients. This leads us to the reason behind this study.

1.5 Purpose of This Study

Even with the advancements in modern treatment of metastatic diseases with IMRT and IGRT, local failures are still common in a large portion of esophageal cancer patients treated with radiation therapy. The treatment strategy for these patients that are non-responders could potentially be altered if these individuals are identified during therapy. This could potentially result in a reduction of tumor failures and overall better long term survival. This work investigates the utility of an F-18 FDG-PET/CT scan acquired during the course of therapy as a predictor of pathological response.

As previously stated, the SUV is a commonly used parameter for the indication of response but currently no clear indicator of what type of SUV metric is best at indicating response has been determined. This study aims to bring clarification to this question by testing the two most common SUV metrics (maximum and mean) along

with several other SUV parameters (described later) to find which is the most accurate at determining the responders from the non-responders of the treatment.

Also, further possibilities of using F-18 FDG PET/CT scans are tested to for the indication of response to treatment besides that of the realm of the SUV metrics. This is done by analyzing textural features of the PET/CT scans that are acquired during the treatment process. These textural features will be tested to see which, if any at all, could be used to determine response to treatment. These features will also be compared to the results of the SUV analysis to see which of the two methods could better be used as a predictor.

Besides comparing the two different ways of measuring the response of the treatment, this study aims to find a common ground between both of the measurements. This is done by trying to utilize both sets of data together to see if they can result in an even better model for the prediction of the response to treatment. With these predictions, the treatment of patients could be altered to better suit them in the hopes of reducing local tumor failures. Overall, this studies goal is to fully test F-18 FDG PET/CT scans for the utilization of a measurement of response to radiotherapy in esophageal cancer patients and by doing so possibly improve the treatment techniques that are currently in place.

2. Materials and Methods

2.1 Patient Background

A total of 28 patients diagnosed with esophageal cancer and undergoing neoadjuvant chemo-radiation were enrolled into an IRB approved clinical trial at Duke University Medical Center. Of these patients, 18 were diagnosed with Adenocarcinoma with the remaining patients having a diagnosis of Squamous Cell Carcinoma. The majority of the patients (16 of the 28) were assigned a clinical tumor stage of T3 and the rest having a clinical tumor stage of T2. All but five of the patients were male and the average age of the patients was 66 years old. The typical treatment plan for these patients consisted of receiving 50.4 Gy over 28 fractions (1.8 Gy/fraction).

2.2 Acquisition of the PET/CT Scans

Prior to treatment, the patients were given F-18 flourodeoxyglucose (FDG) as a contrast agent for the PET scan and a PET/CT scan was administered. A physician then contoured the gross tumor volume (GTV) on the CT image so that treatment planning could begin. After receiving approximately 32 gray (Gy) of the radiation treatment, the patients were once again given the F-18 FDG and a second set of PET/CT images were administered and treatment continued.

2.3 Determination of a New GTV

A new GTV was needed for the study since the size and method for drawing the GTV is dependent on the physician's training and experience level. If two different physicians were to draw the GTV for the same patient the resulting contour may be different as one physician might include certain details that the other does not. To bring more consistency to the GTV's between patients, a new GTV was created from the PET/CT images that were taken prior to the treatment by using a threshold of >40% of the maximum standardized uptake value (SUV) in the physician drawn GTV. Studies have shown that this >40% of the maximum SUV threshold value not only brings consistency but also better represents the actual size of the tumor [8][9]. The standardized uptake value is defined as the ratio of the concentration of activity in a smaller region (i.e. tissue) at some time t, to the total injected activity per body weight of the patient. A mathematical representation of the SUV can be found in equation 1 [10]. C (t) represents the concentration of the injected radioisotope, F-18 in this study, as a function of time and A(t) represents the total injected activity. The unit for the

concentration is typically in MBq/kg while the unit for activity is typically MBq and body weight in kg.

$$SUV = \frac{C(t)}{A(t)/Body\ Weight} \qquad (1)$$

The new GTV was delineated as the GTV_pre_PET. Figure 1 shows how this new contour compares to that of the original physician-drawn GTV.

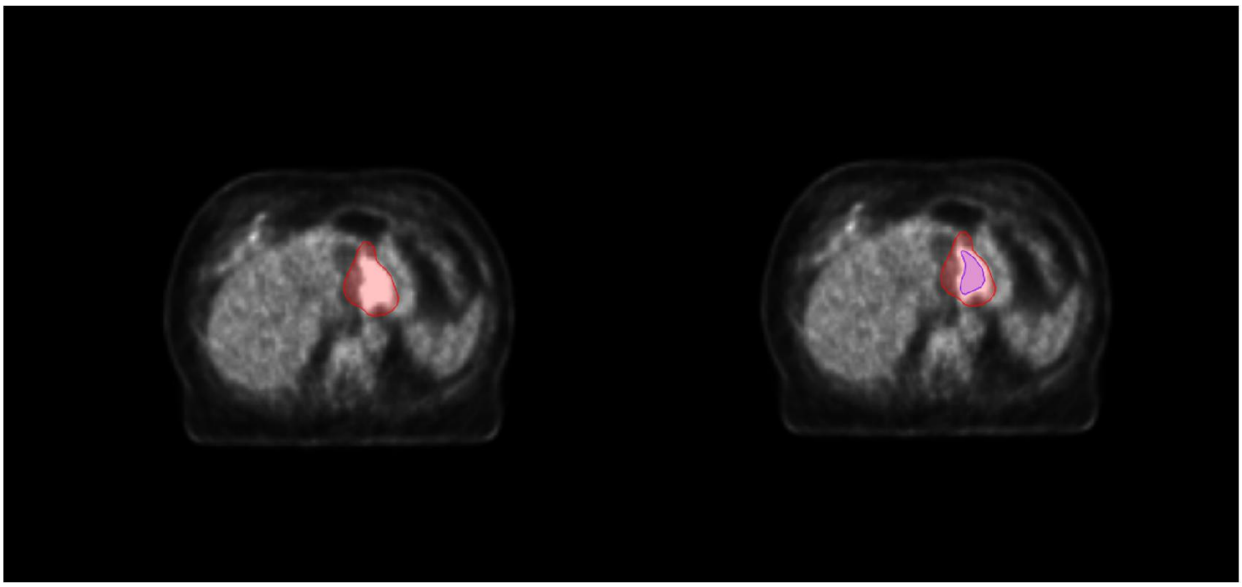


Figure 1: Example of Physician Contoured GTV by Itself (Left) and a GTV Using the 40% Maximum SUV as a Threshold Superimposed on the Original GTV (Right).

The images in Figure 1 are transverse slices of a PET scan from a patient. The tumor can be seen in the images as the brighter part inside the GTV contours. The left image shows just the GTV contoured by a physician in red while the right image has the

GTV_pre_PET contour added by using the 40% of the maximum SUV as a threshold and is colored purple. The GTV_pre_PET contour is superimposed on top of the original GTV in the right image to show the difference in size between the two contours.

2.4 Transfer of the GTV_pre_PET between Images via Image Registration

Now that a new GTV was created from the original contour, it needed to be transferred to the PET image that was taken mid-treatment. To accomplish this, the pre-treatment CT image was registered to the intra-treatment CT scan. This was done by using the registration algorithms in VelocityAI (need to add TM or R symbol, company name, location of company) software. Registration was necssary since the two sets of images were taken at different times, the patient position as well as certain anatomy could be different. A patient might lose weight during the course of the treatment or the patient's soft tissue anatomy (i.e. intestines, bladder, etc.) might shift resulting in the displacement of the tumor or other organs. Registering the images together finds similarities between the images so that information in the images can be shared. First, rigid registration was performed on the images to align the bony structures. Rigid registration looks for similarities between solid structures of the patient's anatomy that typically does not change too much between images.

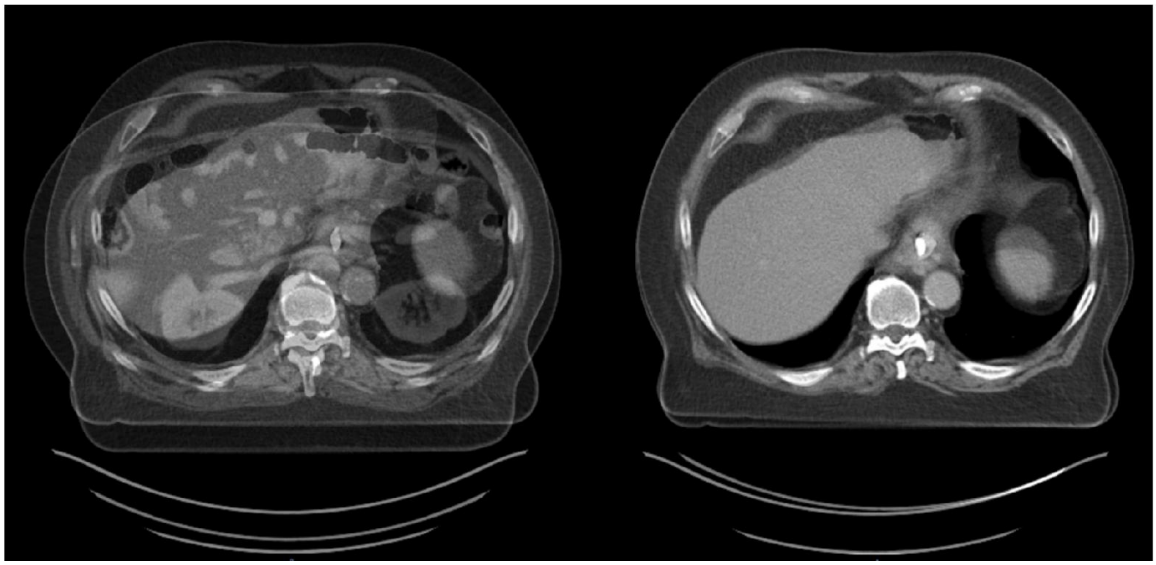


Figure 2: Visualization of No Registration (Left) and After Rigid Registration (Right).

Figure 2 shows the two sets of CT images transposed on top of one another before the rigid registration (Left) and then the resulting image after the rigid registration was completed (Right). Although the bony anatomies are registered, there is still some differences between the two images. This could be seen by how some of the soft tissues in the images look to not be aligned or blurry and therefore data cannot shared between them since they are not in the same coordinate plane. In order to place the two images on the same coordinate plane deformable registration was then performed. Deformable registration compares various grids and nodes between images to perform interpolations to match information. There are currently several different algorithms to perform deformable registration. The deformable registration was performed utilizing VelocityAI's software. Their algorithm is based on Mattes

formulation of mutual information known as the B-Spline model. The deformable registration strategy that was used was the deformable multipass since it is the method that is recommended by Velocity Medical Systems for clinical use [11].

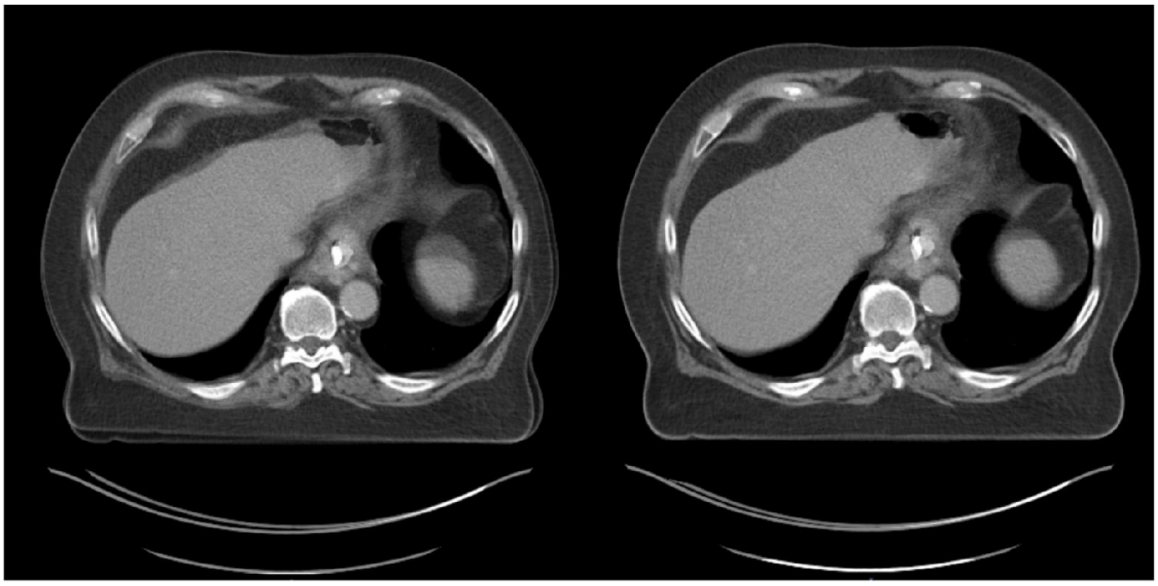


Figure 3: Visualization of the Difference between Rigid Registration (Left) and Deformable Registration (Right)

Figure 3 shows the images after deformable registration was performed. It should be noted that although the image now look like everything lines up perfectly, the one image has been deformed in order to be in the same plane as the other. To try and minimalize the deformation of the image a region of interest (ROI) was created in only a small section of the patient's body that contained the tumor. Since the images were now in the same coordinate plane, data and information could be shared and transferred from the one CT image to the next. The GTV_pre_PET was moved from the pre-treatment PET image to the pre-treatment CT scan. There was no need to register these

two images together since the two images were taken essentially at the same time while the patient was in the same position on the PET/CT scanner. Now that the contour was in the pre-treatment CT image, it was transferred to the now registered intra-treatment CT image. Finally the contour was moved from the intra-treatment CT scan to the intra-treatment PET scan. The contour that was now on the intra-treatment CT scan was delineated as the GTV_intra_PET.

2.5 Analysis of Changes in the Two Sets of Images

Now that the GTV was on both sets of images, analysis could then be performed to test the differences between the two scans. In order to test the differences, metrics were needed to be collected and analyzed that were on both images. Since both the pre and intra-treatment PET scans contain SUV values, they could be used to test the differences between them. As mentioned in the introduction, the most used SUV metrics to analyze response are the fractional decrease in the maximum SUV and the mean SUV. These values were tested along with several new metrics in order to find which, if any, would prove to be the best at indicating response to the treatment. The GTV_pre_PET and the GTV_intra_PET were used as an ROI to find the SUV values that was contained inside of them. The new metrics that were tested were the fractional decreases in the SUV to the highest 10, 20, 30, 40, 60, 80, and 90 percent volume. From these metrics, the most significant values would then be combined to also test for significance. The combination of these metrics is found by simply taking the average of the fractional

decrease of the metrics being combined. Complementary to SUV analysis, another analysis was used to see if it could be used as an indication of response. This was texture analysis of the images. Texture analysis, finds the images texture based on the spatial relationships and arrangements between the image's pixels. This can be done by looking at the changes in the intensity patterns or gray tones. To find the texture of the image, a neighborhood gray tone dimension matrix (NGTDM) was used as described by Amadasun et al. This matrix was then used to find several texture features associated with the images. These texture features were coarseness, contrast, busyness, complexity, and texture strength.

2.5.1 Calculation of the SUV Metrics

The calculation for the SUV was previously described in equation 1 and was done by the VelocityAI software. To calculate the various fractional decease in the various SUV metrics to the highest percent volume, an SUV volume histogram was needed. An SUV volume histogram is a graph that shows how much SUV occupies a certain amount of volume. To create this graph, the SUV by volume (SUVbw) and their corresponding volumes were exported from VelocityAI. The SUVbw exported from VelocityAI was in increasing order. Each increase represented the SUV value occupying a smaller volume. A type of cumulative frequency (CF) of the volume (CC) was then calculated using equation 2. The cumulative frequency is the total of a frequency and all the frequencies that have occurred in the distribution. In this case it is the total of the

volume and all the volumes that have occurred. This cumulative frequency was then normalized to the maximum value in the cumulative frequency so that each value represented a fraction of the volume as it accumulated. This is represented by equation 3 where CFnorm represents the normalized cumulative frequency. This normalized cumulative frequency needed to then be subtracted from one so that they represented the highest percent volume shown in equation 4 with %CC representing the percent volume calculated.

$$CF = \sum_1^i CC_i \quad (2)$$

$$CFnorm = \frac{CF}{CF_{max}} \quad (3)$$

$$\%CC = 1 - CFnorm \quad (4)$$

Now that both the percent volumes and the SUVbw values were obtained, the SUV volume histogram was able to be generated. The decreasing percent volume was plotted against the increasing SUVbw to obtain the histogram. Figure 4 shows an example of a generated SUV volume histogram for both the pre and intra values from a patient.

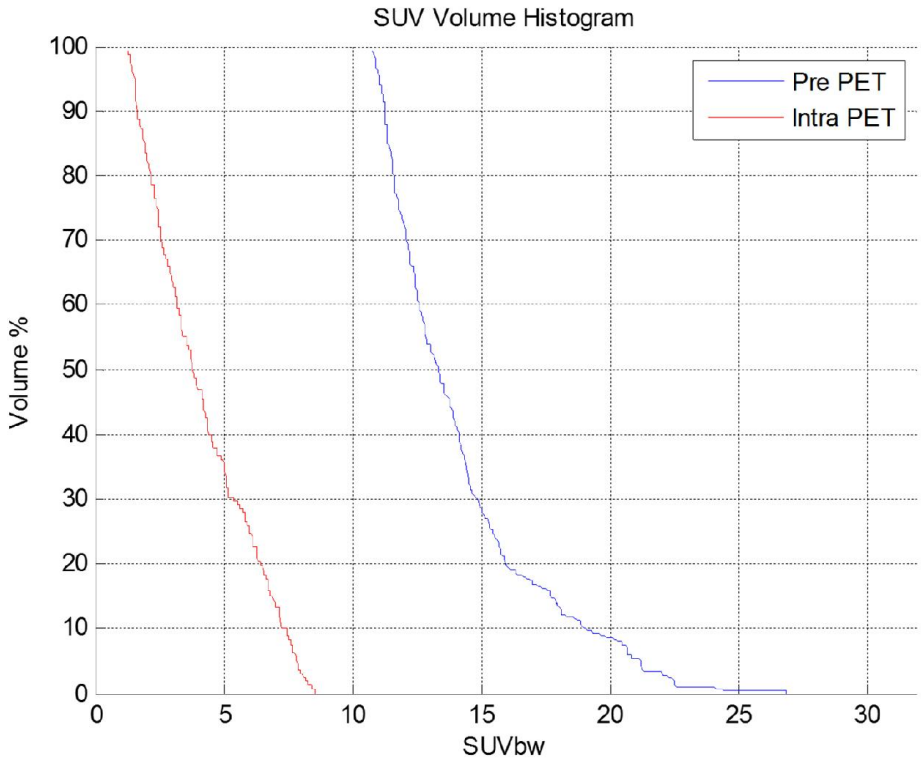


Figure 4: SUV Volume Histogram with Both Pre PET (Blue) and Intra PET (Red).

It can be seen that as the SUVbw increases, the percent that it occupies decreases. Also the SUVbw for the pre-treatment PET scans tend to be higher than that of the values for the intra-treatment PET scan. To find the desired SUV metrics to the highest percent volume, the graph could be used to go to the percent volume on the y- axis and then traced across the x-axis to both the pre and intra-PET plots and then record both respective SUVbw values. Figure 5 shows an example of this for the SUV to the highest 20 percent volume (SUV20).

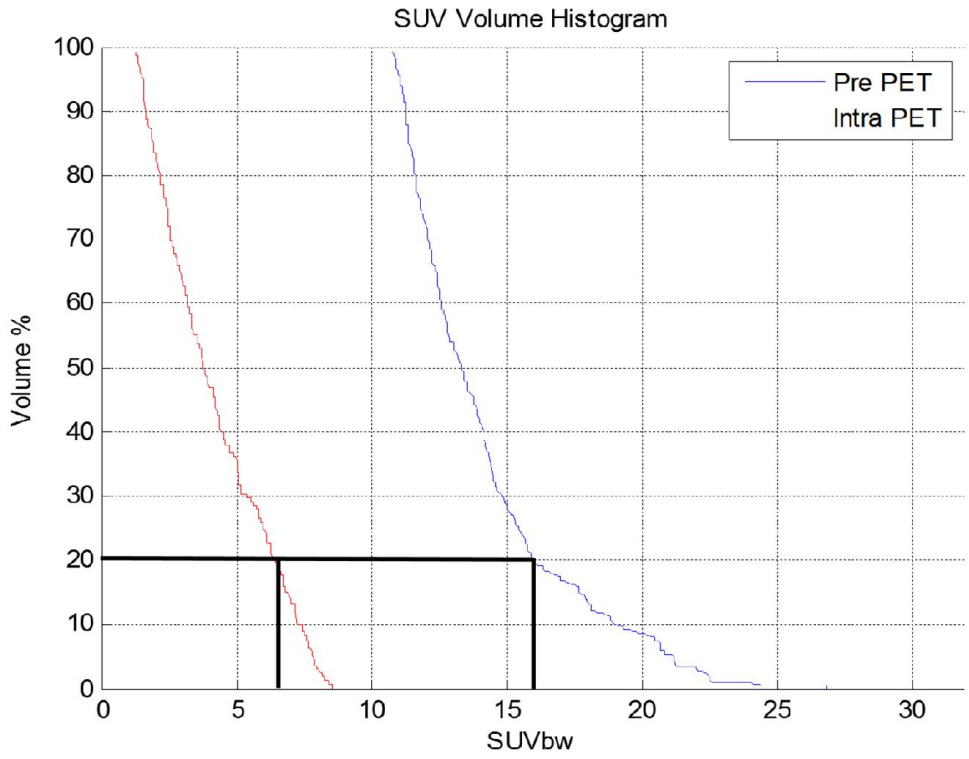


Figure 5: SUV Volume Histogram Showing How to Acquire the SUV Data for SUV20.

Using the SUV volume histogram, each of the SUV metrics to the highest percent volume could be could be found. An in house Matlab program was written to find these values based on the data that was used to generate the histograms and were recorded for each patient. The SUVmax value was found by simply recording what the maximum SUVbw value was from the data. The SUVmean value, however, required a small calculation to find it. Equation 5 shows how this calculation was performed.

$$SUVmean = \frac{\sum_1^i (SUVbw_i \times CC_i)}{\sum CC} \qquad (5)$$

In equation 5, each SUVbw value is multiplied by its corresponding volume value that was originally exported from VelocityAI and then summed. This is then divided by the sum of all of the volume data. Once again, an in house MATLAB program was written to calculate the SUVmean and find the SUVmax value.

2.5.2 Calculation of the Texture Features

Texture can be simply defined as how the most fundamental parts of a material are arranged. In digital imaging, a texture can be found from looking at the differences in the intensities acquired from a pixel's gray tone and the gray tones of its closest neighbors. The development of the different texture features can be found by calculating a one dimensional matrix for a particular image. This matrix is known as a neighborhood gray tone difference matrix (NGTDM) and its computation is described in equations 6 and 7 as reported in a study by Amadasun et al [12].

$$\bar{A}_i = \bar{A}(k,l) = \frac{1}{W-1}\left[\sum_{m=-d}^{d}\sum_{n=-d}^{d} f(k+m, l+n)\right] \qquad (m,n) \neq (0,0) \qquad (6)$$

$$s(i) = \sum |i - \bar{A}_i|, \quad for\ i \in N_i\ if\ N_i \neq 0$$

$$s(i) = 0, \quad otherwise \qquad (7)$$

Equation 6 is the calculation of the average gray tone over a neighborhood centered at position (k,l). The neighborhood is specified by a size d and the term $W= (2d+1)^2$. In equation 7, this average gray tone is subtracted from the gray tone value in question (i) to calculate the ith entry of the NGTDM. N_i represents the set of pixels that have gray

tone level *i* in the neighborhood. Using the NGTDM, a variety of textural features can be determined. These textual features include coarseness, contrast, busyness, complexity, and texture strength. A description of these features and their calculation can be found in the following subsections. An in house MATLAB program was used to calculate the NGTDM and each of the texture features for both the pre-treatment PET scans and intra-treatment PET scans.

2.5.2.1 Coarseness

Coarseness is the most fundamental feature of texture and in its most narrow sense it implies texture. When a texture is coarse, the most basic patterns that make up the texture are large and as a result there tends to be a large degree of uniformity in intensity and the spatial differences in these intensities are subtle. This leads to there being smaller differences between a gray tone value of a pixel and the average gray tone of its neighboring pixels. The summation of these small differences using every pixel in the entire image will result in the spatial change of the intensity and inversely the coarseness. The calculation for coarseness for an NxN image is described in equation 8 [12].

$$f_{cos} = \left[\epsilon + \sum_{i=0}^{G_h} p_i s(i)\right]^{-1} \qquad (8)$$

$$p_i = \frac{N_i}{n^2}, \quad where\ n = N - 2d. \quad (9)$$

G_h in equation 8 is the largest gray tone value that is in the image and ϵ is simply a small number that prevents the outcome from becoming infinite. p_i is the probability of occurrence of gray tone i in the image which can be computed utilizing equation 9.

2.5.2.2 Contrast

During analysis an image, if there are areas that have a clearly visible difference in the intensity levels then the image is said to be one with high contrast. If there is a large difference in the intensities between the pixels then there is also a large difference between neighboring gray tone level. Therefore, when there is a large difference in the gray tone level of a pixel and its neighbors it is said that there is high contrast. The spatial frequency of the differences in intensity also affects the contrast (e.g. a small checkerboard will have a higher contrast than a coarser checkerboard even though the gray tone scale is the same). Taking these into account, a calculation of contrast is able to be made and is shown in equation 10 [12].

$$f_{con} = \left[\frac{1}{N_g(N_g-1)}\sum_{i=0}^{G_h}\sum_{j=0}^{G_h} p_i p_j (i-j)^2\right]\left[\frac{1}{n^2}\sum_{i=0}^{G_h} s(i)\right] \qquad (10)$$

$$N_g = \sum_{i=0}^{G_h} Q_i\,, \qquad where\ Q_i = \begin{cases} 1, & if\ p_i \neq 0 \\ 0, & otherwise \end{cases} \qquad (11)$$

As in equation 8, p_i and p_j represent the probability of occurrence for gray tone value. $N_g$is the total number of different gray tone levels that are in the image and can be found using equation 11. It can be seen that $f_{con}$is the product of two terms with the first

term representing the average weighted squared difference between different grey tone values and the second term being an average difference of the NGTDM.

2.5.2.3 Busyness

A texture of an image is considered to be busy if there a rapid changes in the intensity between one pixel and its neighbors which means that the spatial frequency in the differences of intensity are high. Even though this spatial frequency reveals how much busyness there is, the magnitude of the changes in intensity is dependent on the range of the gray tone scale and therefore can be related to the contrast. A suppression in the contrast information about the spatial rate of change in the intensity can indicate the level of busyness in the image. This leads to the calculation of busyness described in equation 12 [12].

$$f_{bus} = \left[\sum_{i=0}^{G_h} p_i s(i)\right] \Big/ \left[\sum_{i=0}^{G_h}\sum_{j=0}^{G_h} ip_i - jp_j\right], \quad p_i \neq 0, p_j \neq 0 \qquad (12)$$

Here the numerator is the measure of the spatial rate of change in the intensity and the denominator is the summation of the magnitude of the differences in the different gray tone values. The denominator is the suppression of the information about the spatial rate of change in the intensity in the contrast and emphasizes the frequency of the differences in intensity values.

2.5.2.4 Complexity

The complexity of a texture refers to how much visual information content in the image there is. A texture would be considered complex if there is a lot of visual information content available which happens when there is a lot of basic patterns in the image and even more when these basic patterns have a varying degree of average intensities. An image that has a lot of sharp edges or lines would be an example of having a texture that is complex. When there is a lot of spatial changes in the intensity in the image it is more likely to be complex than an image that has a more uniform distribution of intensities. A lot of spatial changes in the intensities would result in a large number of different gray tone values but would decrease the probability of occurrence of each individual value. This leads to the size of the basic patterns in the image and the probability of occurrence having an inverse relationship with the complexity of the texture. The calculation of the complexity of a texture is described in equation 13 [12].

$$f_{com} = \sum_{i=0}^{G_h} \sum_{j=0}^{G_h} \left\{ (|i-j|) / \left(n^2 (p_i + p_j) \right) \right\} \{ p_i s(i) + p_j s(i) \}, \quad p_i \neq 0, p_j \neq 0 \qquad (13)$$

The term $\left(n^2(p_i + p_j)\right)$ in equation 13 represents the inverse relationship between complexity and the sizes of the basic patterns and probability of occurrence of the values in the image. f_{com} is the sum of the normalized differences in the intensity values weighted by the NGTDM and taken in pairs. The absolute difference in the gray tone values represents the influence in the variations of contrast on the complexity of the image.

2.5.2.5 Texture Strength

A strong texture is one where the basic patterns that it is made up of are clearly definable and visible. Although a texture may generally look attractive and present a high level of visual feel, the ease at which distinctions can be made between the basic patterns is dependent on the actual size of these basic patterns and the change in their average intensities. One may be able to tell a difference between large basic patterns even though there is not a large difference between the average intensities however for this distinction to be made when there are smaller basic patterns there has to be a large difference between the average intensities. As a result of this, the strength of a texture can be correlated with both coarseness and contrast. Equation 14 shows the computation of the strength of a texture [12].

$$f_{str} = \left[\sum_{i=0}^{G_h}\sum_{j=0}^{G_h}(p_i + p_j)(i-j)^2\right] \Big/ \left[\epsilon + \sum_{i=0}^{G_h} s(i)\right], \qquad p_i \neq 0, p_j \neq 0. \qquad (14)$$

The numerator in the equation stressed the differences in the intensities and can represent the intensity differences between the basic patterns of the image. The denominator can reveal information regarding the size of these basic patterns and is just the sum of the NGTDM for all of the gray tone values. The denominator in the equation for texture strength is the same denominator in the calculation of the coarseness in a texture. Therefore, texture strength is proportional to the coarseness of a texture. The entire equation emphasizes the boldness of the basic patterns and hence a large resulting value would indicate a strong texture.

2.6 Recording Responders and Non-Responders

The SUV metrics and the texture features are used to indicate whether or not the patient was responding to treatment. In order to do this there needed to be a way to tell which of the patients really was responding to the treatment. To achieve this, the patients underwent a surgical biopsy on the tumor site. A physician then created a pathology report based on this biopsy and gave the tumor a treatment effect grade. This treatment effect grade ranged from 0 to 3 with 0 indicating that there no or little trace of the tumor left and 3 indicating that the tumor was still very present. To classify how the patients responded to the treatment, the patients whose treatment effect grade was 0 or 1 were deemed to be a responder while those whose treatment effect grade was 2 or 3 were deemed to be non-responders. Not all of the patients who entered the trial underwent this surgical biopsy procedure. A total of 20 patients in the trial underwent surgery and were then classified.

2.7 Analysis of the SUV Metrics and Texture Features

To analyze the SUV metrics and the texture features, the fractional difference between the data for the pre and intra-treatment PET scans was measured. This fractional difference was found by subtracting the pre-treatment PET data from the intra-treatment data and then dividing it by the pre-treatment data.

2.7.1 Wilcoxon Rank Sum Test on the Data

The fractional difference was then compared to the responders and non-responders by using a Wilcoxon rank-sum test (also called the Mann-Whitney U test). This test takes two distributions, in this case the differences in the data for either a responder or non-responder, and places them in an ascending order. Once in ascending order, a rank is assigned to the corresponding value dependent upon where it falls in the order. If there are values that are equal to each other, the rank that is assigned to them is the mean of their position in the order. Once a rank is assigned to each value, the data is placed back in their respectful distributions with their ranks accompanying them. The ranks are then summed to create a test value. The smaller of the two distribution's test value is used to test against the null hypothesis. The null hypothesis in the study was that the data could not indicate a difference between responders and non-responders. The alternative hypothesis is that the two sets of data could indicate a difference from the responders and non-responders. The null hypothesis was said to be rejected if the resulting p-value from the Wilcoxon rank sum test was less than 0.05. The Wilcoxon rank sum test was also performed on pre-treatment and intra-treatment texture features individually [13] [14].

2.7.2 Finding Sensitivity, Specificity, and a Threshold

After the Wilcoxon rank sum test was performed, a calculation to find the sensitivity (true positive rate) and specificity (true negative rate) was done and used to

find a threshold value for the difference between the data in the pre and intra-treatment PET scans that would accurately identify a responder or non-responder. Sensitivity measures the proportion of the true positives in the data that were correctly identified as such. When the data correctly identifies a positive it is known as a true positive. Sensitivity measures the proportion of the data that correctly identifies a negative result. When this occurs, the data point that correctly identifies a negative is called a true negative. Sometimes the data will identify a positive when the actual outcome is a negative or identify a negative when the outcome is positive. These situations are known as being false positives and false negatives. When a false positive occurs, it is known as a Type I error while a false negative is a Type II error. Equations 15 and 16 show the formula for calculating sensitivity and specificity. Values of sensitivity and specificity range between 0 and 1 with values closer to 1 being optimal since this indicates that it was able to correctly identify both true positives and true negatives.

$$Sensitivity = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives} \quad (15)$$

$$Specificity = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives +\ Number\ of\ False\ Positives} \quad (16)$$

To find the threshold value, the fractional difference data was first sorted into ascending order. Next, a series of values were tested for sensitivity and specificity based on the order of the data. The values were selected by taking a data point and then averaging it with the next data point in the order until all of the data was including in

calculating a value. For example, if the ascending data was 1, 2, 3, and 4 then the resulting threshold values that were tested would be 1, 2.5, and 3.5 respectively. The overall threshold value was said to be the one that had both a large sensitivity and specificity value.

2.7.3 Receiver Operating Characteristic Curve

Now that the sensitivity and specificity values were calculated, a receiver operating characteristic (ROC) curve was able to be produced. An ROC curve is a plot of the true positive rate (i.e. sensitivity) versus the false positive rate (i.e. 1-specificity). ROC curves are commonly used in diagnostic purposes to tell if a result had been correctly classified. In this study, it was used to see if the analysis was correctly able to identify if a patient was a responder or a non-responder. In ROC analysis there is typically a tradeoff between sensitivity and specificity. This tradeoff is typically dependent on what the threshold value is for determining a result. If the threshold value is high, this will lead to a generally high sensitivity and a low specificity. Conversely, if the threshold value is chosen to be smaller, this will lead low sensitivity and a high specificity. When the sensitivity is high, the model is more likely to miss more negative cases while a low sensitivity will have a tendency to miss more positive cases. As previously mentioned, an ROC curve is a plot of the sensitivity versus 1-specificity. Generally a good resulting curve will be one that trends toward the upper left hand corner. A perfectly accurate curve is essentially two line segments that would take up

the entire left hand corner. A plot that would be just a 45 degree line would mean that the model was making a blind guess [15]. The area under the ROC curve (AUC) was then computed for the collected data. The AUC is used as an indicator for the diagnostic performance of the model and is the probability that the model will correctly identify a positive result when there is a case that is positive and a case that is negative. Both the ROC curves and the AUC were computed using an in house MATLAB program.

2.7.4 Logistic Regression Analysis

Since this study used to completely unrelated distributions of data (SUV analysis and texture feature analysis.), a way to use both of these distributions together to find an indication of response was investigated. To do this a type of linear regression was used to include both the most significant SUV terms and most significant texture feature. The type of linear regression that was utilized was logistic regression. In logistic regression analysis, there is a variable that is binary. That is, a variable that is either "yes" or "no" or a "1" or "0". In this study that variable is the responder or non-responder. The question that is being asked in the analysis is if you can predict the probability of the outcome of this variable from an input variable X. To answer this, a type of linear regression is used and set equal to a logit function as shown in equation 17. In linear regression, data is fit to a linear equation by multiplying the input variable by a constant (α in equation 17). The difference between standard linear regression and logistic regression is how these constants are calculated. Standard linear regression uses a least-

squares method of calculating the constants but since logistic regression is predicting probabilities, a maximum-likelihood method is used. This method finds values of the parameters that would most likely give you the observed result. In equation 17, p(x) represents the probability as a function of the input variable X. The linear equation can have multiple inputs with a new constant being found each time a new variable is added. Note that logistic regression does not directly find the probability but instead gives a logit that when plotted yields an inversed sigmoid shape. Solving for p(x) gives the desired equation for the probability of the outcome as can be seen in equation 18.

$$\log\frac{p(x)}{1-p(x)} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots \alpha_i X_i \qquad (17)$$

$$p(x) = \frac{e^{\alpha_0+\alpha_1 X_1+\alpha_2 X_2 \ldots \alpha_i X_i}}{1+e^{\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\ldots \alpha_i X_i}} \qquad (18)$$

Since the outcome of equation 18 is a probability, the resulting values will have a range from 0 to 1. When plotted, a sigmoid shaped curve is seen. Since the goal of linear regression is to predict the binary variable, it is conventional to state that a $p \geq 0.50$ results in a binary value of 1 while a $p < 0.50$ would result in the binary value of 0. In this study, if the outcome resulted in a 1 then it was said to predict a non-responder and if the outcome was a 0 then it would predict a responder. The input variables used were the SUV metrics and textural features that were found to be the most statistically significant. A Wilcoxon rank-sum test was performed on the data to make sure that it would be able to distinguish between responders and non-responders. Sensitivity and

specificity analysis was also performed to test what threshold probability value would best be used to predict if the outcome was a responder or non-responder.

2.7.5 Cross Validation of the Results

Cross validation is used to assess how a model will perform on an independent data set. It is particularly useful for models where a prediction of the outcome is trying to be made and can tell whether or not the model can make accurate predictions. In k-fold cross validation, the data set is split into smaller groups called folds. The letter k in k-fold cross validation is an indicator for how many folds you split the data into. Each fold can be thought of as an independent data set that will be tested on the model. One of the folds is left out of the calculation of the model and then used to test the validity of it. The fold that is left out is called the test set while the folds that are used in the calculation of the model is known as the training set. The process of leaving out a data set and then using it on a calculated model is repeated until each fold has been used as the test set. Each time the test set is used, an estimation of the accuracy can be made. The overall accuracy of the model can be calculated from each of the accuracies from the individual tests.

In this study, a 3-fold cross validation was performed on the fractional difference tests. As discussed in section 2.7.2, a threshold value was calculated by evaluating a training set while leaving out a test set. This was repeated until each of the 3 folds was used as a test set resulting in 3 threshold values from the training sets. The

threshold values were used on their respective test set for a prediction of treatment response. Since the actual response to these patients in the test sets were known, a count of the true positives, false positives, true negatives, and false negatives was able to be made. This allowed for a calculation of sensitivity and specificity of the cross validation scenario showing the accuracy of the model.

Since the logistic regression fit was being used to create a new model to predict response. A more robust type of cross validation was needed to be performed. This was done by randomly splitting the data. Since there was 9 responders and 11 non-responders, 6 responders and 7 non-responders were chosen to make a training group of 12. This means 84 different combinations of 6 different responders and 330 different combinations of different non-responders that could be used. This leads to a possible 27,720 different combinations of 12 that could be used as training groups. For simplicity, 2000 different training data sets were used to create logistic regression fit equations. These equations were then tested on their corresponding left out test data sets. The average threshold, sensitivity, and specificity was then calculated along with their standard deviation.

2.8 GTV using 20% as a Threshold

Some studies have shown that using a threshold of >20% of the maximum SUV value is the best indication of the actual tumor size. [18] A new GTV was created using this >20% of the maximum SUV value the same way that the GTV for the 40% threshold

was. The contour was then transferred from the pre-treatment PET/CT scans to the intra-treatment PET/CT scans using rigid and deformable registration as previously discussed in section 2.4. Once the contour was on both sets of PET scans the fractional decrease in the same SUV metrics as the 40% threshold were calculated and recorded. Analysis was then performed using a Wilcoxon rank sum test and the ROC curve. Once a metric was deemed significant, cross validation was performed to show the robustness of the data. Texture analysis was not performed for the threshold of >20% of the maximum SUV value.

3. Results

Of the 28 patients that entered the trial, only 20 had undergone surgery to biopsy the tumor. After the physician determined the treatment effect on the patients it was found that 9 of the 20 patients met the criteria of being a responder and 11 of the 20 patients met the criteria for being non-responders. These 20 patients were used in the finding of the following results.

3.1 Results for SUV Metrics

Table 1: Results from Wilcoxon Rank-Sum Test (p-value) and AUC of the calculated ROC curve on SUV Metrics with >40% threshold.

SUV Metric	p-value	AUC
Max	0.095	0.7272
Mean	0.081	0.7373
Highest 10%	0.040	0.7778
Highest 20%	0.040	0.7778
Highest 30%	0.040	0.7778
Highest 40%	0.129	0.7578
Highest 60%	0.111	0.8283
Highest 80%	0.129	0.8687
Highest 90%	0.171	0.8889
Combination of 10% and 20% Metrics	0.033	0.7879
Combination of 20% and 30% Metrics	0.033	0.7879

Table 1 shows the resulting p-values from performing a Wilcoxon Rank-Sum test and also the calculated AUC's from the ROC curve analysis on the SUV metrics. Several of the metrics were considered statistically significant based on a resulting p-values that were less than 0.05. These metrics were the SUV to highest 10%, 20%, and 30% volumes, along with several combination terms. These combination terms were created to see if a combination of different statistically significant values would result in a more significant result. To find the combination metrics, the average of the fractional difference from the metrics that were being combined was taken. The combination terms that were tested were a combination of the highest 10% and 20% metrics and a combination of the highest 20% and 30% metrics.

As can be seen from the table, the most statistically significant results were the two combination terms. These resulted in lowest p-values of all the different SUV metrics while also maintaining a large AUC. The p-value for both the combination of the 10% and 20% metrics and the combination of the 20% and 30% metrics was 0.033 and the AUC for both was 0.7879. As previously mentioned, a large AUC shows that there is a large probability of indicating a correct response. The two terms that were previously most commonly used for indicating response (SUVmax and SUVmean) both had large AUC's with 0.7272 and 0.7373 respectively, but were deemed statistically insignificant since their p-values failed to meet the less than 0.05 criteria. Some trends in the data in Table 1 can also be noted. Excluding the combination terms, the AUC's tended to

increase with increasing percent volume starting at 0.7272 with the SUVmax metric and peaking with 0.8889 with the SUV to the highest 90% metric. While this would normally be a desired result, the p-value also tended to increase with the percent volume as well excluding the SUVmax metric.

3.1.1 Threshold, Sensitivity, and Specificity of Significant SUV Metrics

Table 2: Threshold, Sensitivity, and Specificity for Significant SUV Metrics with >40% threshold.

SUV METRIC	THRESHOLD	SENSITIVITY	SPECIFICITY
SUV 10%	0.4665	0.7778	0.7273
SUV 20%	0.4592	0.7778	0.7273
SUV 30%	0.4326	0.7778	0.7273
COMBINATION OF 10% AND 20%	0.4319	0.7778	0.7273
COMBINATION OF 20% AND 30%	0.4369	0.7778	0.7273

Table 2 shows the results from finding the threshold value of the fractional difference and its ability to determine responders (sensitivity) and non-responders for the SUV metrics that were found to be significant from section 3.1 above. It can be seen that all of the significant metrics had high sensitivities and specificities for their threshold level. In fact, all of the metrics had the same sensitivity and specificity with 0.7778 and 0.7273 respectively. Although these values were the same the threshold value

used to determine them were different between the different metrics. SUV to the highest 10% yielded the largest threshold value with 0.4665 while the combination metric of the highest 10% and 20% had the smallest threshold value of 0.4319. The threshold values were very similar to each other with the difference between the largest value and smallest value being 0.0346 and the average threshold of all of the significant metrics being 0.4454.

Figure 6 shows the plotted ROC curves from these significant SUV metrics. The red marker in the figure indicates where the threshold value is able to distinguish the sensitivity and specificity for the different metrics. Notice how all of the plots bend towards the upper left hand corner. As previously discussed in section 2.7.3, this is the desired result of a ROC curve.

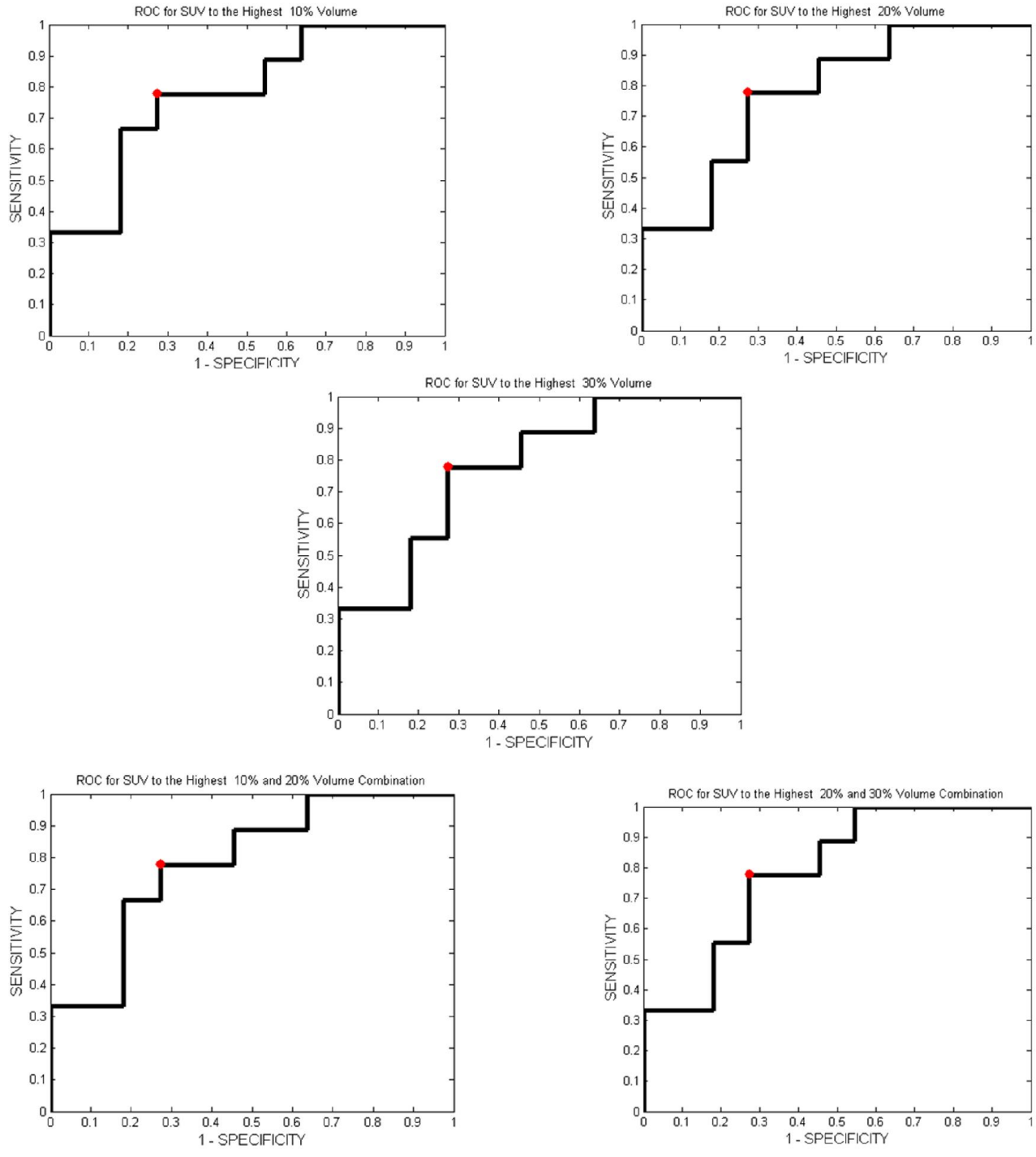


Figure 6: ROC Curves for Significant SUV Metrics Using a >40% Threshold.

Of the five plots, it appears that the combination term of SUV to highest 10% and 20% bends more significantly to the upper left hand corner than the other plots. The plot for the SUV to the highest 20% and the plot for the SUV to the highest 20% volumes appear to be almost identical and look to be the plots that bend the least to the upper left hand corner. None of the plots generated were close to a 45 degree line which would indicate a random guess for distinguishing responders and non-responders.

3.1.2 Cross Validation Results For >40% Threshold

Table 3: Results from 3-Fold Cross Validation of the SUV Metrics Using >40% Threshold.

SUV METRIC	SENSITIVITY	SPECIFICITY
HIGHEST 10%	0.7143	0.6400
HIGHEST 20%	0.7143	0.6400
HIGHEST 30%	0.7143	0.6400
COMBINATION OF 10% AND 20%	0.8333	0.7727
COMBINATION OF 20% AND 30%	0.8333	0.7273

Cross validation of all metrics that were significant yielded results with relatively high specificity and sensitivity. This showed that each model could be used to produce good results on an unknown data sample and thus be accurate. The SUV to the highest

10%, 20%, and 30% all had the same sensitivity and specificity with 0.7143 and 0.6400 respectively. The two combination terms both had sensitivity and specificity results that were higher than that of the individual metrics. The combination of the highest 10% and 20% metrics yielded a sensitivity that was 0.8333 and a specificity that was 0.7727 while the combination of the highest 20% and 30% yielded a sensitivity of 0.8333 and a specificity that was 0.7273.

3.1.3 Best SUV Metric for >40% Threshold

From the results of the Wilcoxon rank-sum test, ROC analysis, and cross validation, one metric seemed to stand out amongst the other significant results. This metric was the combination metric of the 10% and the 20% metric. Not only was it one of the two metrics that yielded the lowest p-value and highest AUC, it also gave the highest sensitivity and specificity after undergoing 3-fold cross validation. Also when observing all of the ROC plots together in figure 6, this metric appeared to give the best desired bend towards the upper left hand corner. Therefore it can be seen that the best SUV metric for use as an indication of response between pre-treatment and intra-treatment PET/CT scans is the average fractional difference of the highest 10% and 20% SUV metrics.

3.2 Results for Textural Features

Table 4: Results from Wilcoxon Rank-Sum Test (p-value) and AUC of the calculated ROC curve on Texture Features.

Texture Feature	p-value	AUC
Contrast	0.171	0.3131
Coarseness	0.081	0.7374
Busyness	0.323	0.3636
Complexity	0.040	0.7778
Texture Strength	0.171	0.3131

Table 4 shows the resulting p-values from the Wilcoxon rank-sum test and AUC for the ROC's generated for the fractional difference in the different textural features. Two of the features, coarseness and complexity, had relatively large AUC's with 0.7374 and 0.7778 respectively but the only feature that met the criteria for being statistically significant was the complexity feature having a p-value of 0.040. Contrast and texture strength yielded similar results with both having p-values of 0.171 and AUC's of 0.3131. All but two of the features had AUC's under the 0.5 mark. This usually means that the values generating the curve should be reversed. For example, if you were using a decrease in the terms to test sensitivity and specificity, you should now use an increase instead and a mirror of the previous curve would be made. This was not investigated here since the Wilcoxon rank-sum test found that these features could not distinguish

between the two distributions of responders and non-responders. Therefore, the only feature that was deemed significant was complexity.

3.2.1 Threshold, Sensitivity, and Specificity of Significant Texture Features

Table 5: Threshold, Sensitivity, and Specificity for Significant Texture Features.

Texture Feature	Threshold	Sensitivity	Specificity
Complexity	0.2050	0.8889	0.7273

Table 5 shows the threshold value and its ability to distinguish between responders (sensitivity) and non-responders (specificity) for the only significant texture feature, complexity. This yielded a threshold of the fractional decrease of 0.2050 and which could distinguish responders with a very high sensitivity of 0.8889 and distinguish non-responders with a high specificity of 0.7272.

Figure 7 shows the plot of the ROC curve for this texture feature, with the red marker indicating where the threshold value is able to distinguish responders and non-responders. The ROC curve can be seen to bend significantly towards the upper left hand corner of the plot as desired and well away from the 50/50 guessing plot of a 45 degree line. This means that the model can distinguish between responders and non-responders very well.

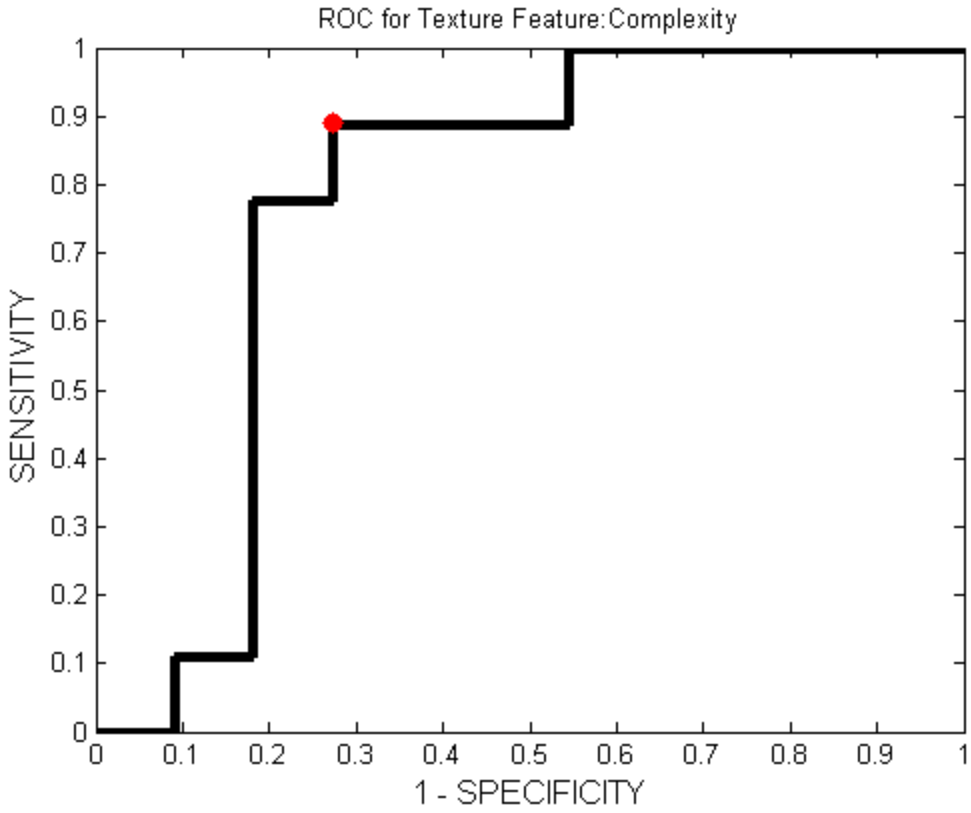


Figure 7: ROC Curve for the Texture Feature Complexity.

3.2.2 Cross Validation of Significant Textural Features

Table 6: Results from 3-Fold Cross Validation of the Texture Feature Complexity.

Texture Feature	Sensitivity	Specificity
Complexity	0.8333	0.7273

Table 6 shows the results from a 3-fold cross validation of using the complexity feature as an indicator for response. Both the sensitivity and specificity were high with the sensitivity being 0.8333 and the specificity being 0.7273. These results show that complexity can be accurately used as an indicator for response since cross validation shows how the model will act when presented with a group of data that was not used in the creation of the model.

3.3 Results from Logistic Regression

Since the combination of the highest 10% and 20% SUV metrics and the texture feature complexity were found to be the two best indicators of response, they were used to find a logistic regression equation. Equation 19 shows the resulting logit function from the logistic regression model.

$$\log\left(\frac{P(X1,X2)}{1-P(X1,X2)}\right) = 2.005 - 3.235X1 - 2.577X2 \qquad (19)$$

Here X1 is the combination of the highest 10% and 20% metric and X2 is the texture feature complexity. P(X1, X2) is the probability that the treatment would result in a patient being a non-responder. The constant terms that were found by using a maximum likelihood method were 2.005 as the y-intercept term, -3.235 as the term that would result in the desired outcome given the combination term, and -2.577 as the term that would have the desired outcome given the complexity feature. It can be seen that equation 19 is in the form of a linear equation as desired. Solving equation 19 for P(X1, X2) results in an equation for the probability as shown in equation 20.

$$P(X1, X2) = \frac{e^{2.005 - 3.235X1 - 2.577X2}}{1 + e^{2.005 - 3.235X1 - 2.577X2}} \qquad (20)$$

Equation 20 yields the desired result of finding an equation that would utilize the most significant results from the analysis of the SUV metrics and also the texture features. Equation 20 outputs a probability that the treatment would result in a non-responder. These resulting probability values were then tested using the Wilcoxon rank-sum test and ROC curve analysis to see how accurately the model was able to distinguish between the distributions of responders from non-responders. Figure 8 shows the probability values plotted in ascending order.

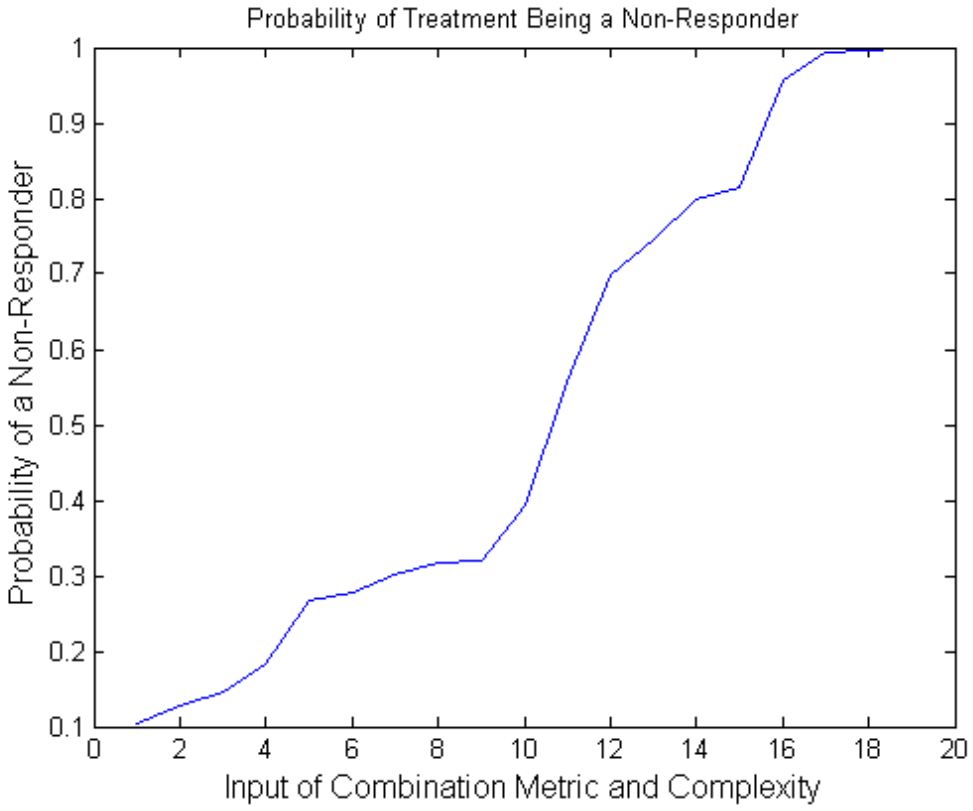


Figure 8: Probability of the Outcome of the Treatment Being a Non-Responder.

It can be seen how the resulting outcomes from solving the logistic regression equation for the probability results in a sigmoid "S" shaped curve. Figure 8 also shows how the outcome of the desired range of values between 0 and 1. These resulting probability values could be used in order to determine whether or not the patient being treated will become either a responder or a non-responder. Table 7. Shows the resulting p-value and AUC of the ROC curve calculated.

Table 7: Results from Wilcoxon Rank-Sum Test and ROC Curve Analysis on Logistic Regression Equation.

Model	p-value	AUC
Logistic Regression	0.004	0.8889

The resulting p-value from the Wilcoxon rank-sum test was 0.004 which is well below the 0.05 parameter that determines significance. This means that equation 20 was able to accurately separate the probability values that were associated with patients that were responders from the probability values of the patients that were non-responders. Also, the resulting AUC of the ROC curve was 0.8889 indicating that the model had a large true positive rate and small false positive rate. This means that the model will predict the correct outcome with an 88.9% probability. Sensitivity and Specificity testing resulted in finding a threshold value for the probability. This threshold indicates that anything above it would be a non-responder while anything below it would be a

responder. Table 8 shows this threshold value along with its corresponding sensitivity and specificity.

Table 8: Threshold, Sensitivity, and Specificity of Logistic Regression Equation.

Model	Threshold	Sensitivity	Specificity
Logistic Regression	0.3186	0.9091	0.7778

The threshold probability was determined to be of 0.3186 which was able to distinguish between responders and non-responders with a sensitivity of 0.9091 and a specificity of 0.7778. This model performed very well being able to accurately predict a true responder with almost 91% accuracy while also being able to predict a true non-responder with an accuracy of 78% of the time. Figure 9 shows the ROC curve for the logistic regression model and as before in the other ROC curves, the red marker indicates the threshold's sensitivity and specificity.

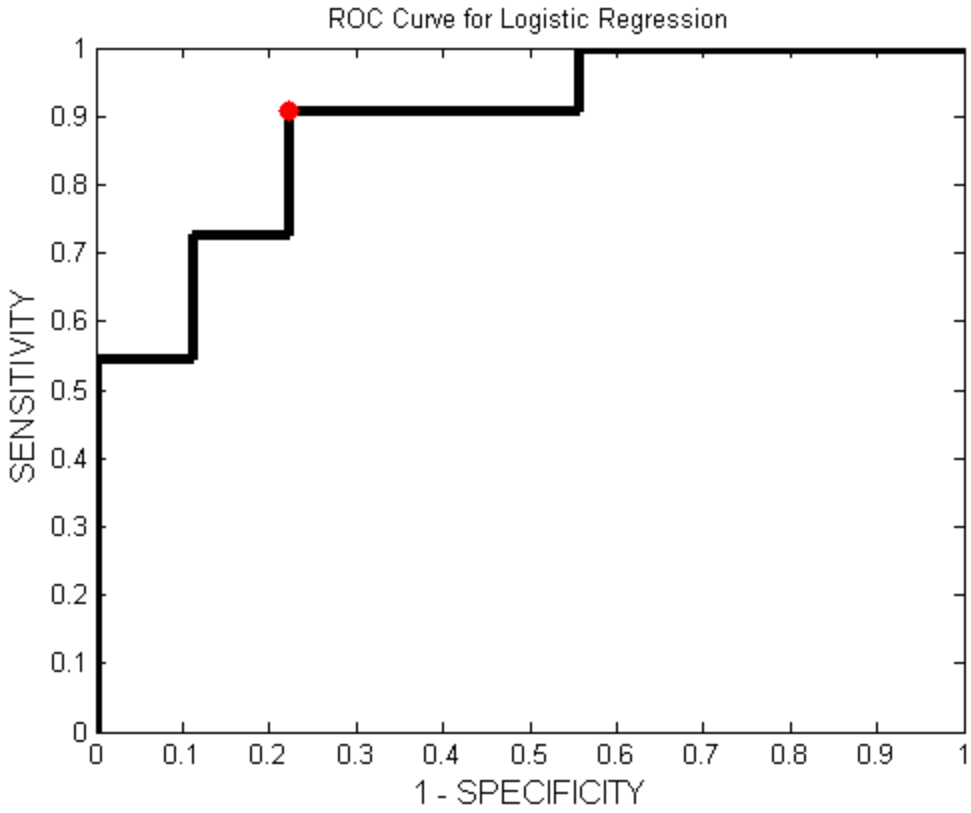


Figure 9: ROC Curve of the Logistic Regression Equation.

It can be seen that the curve in the figure has the desired bend towards the upper left hand corner. Not only does it bend in the right direction, but the bend is very significant and far away from being a 45 degree line which would indicate a guess. This means that the model had a large true positive rate while also having a low false positive rate and distinguish between responders and non-responders.

3.3.1 Cross Validation of the Logistic Regression Analysis

Table 9 shows the results from running cross validation of the logistic regression data that was randomly split into 2000 training groups of 13 and 2000 test groups of 7. The training data was used to create 2000 different logistic regression models and then the 2000 test data sets were used to test their respective model. The average sensitivity and average specificity values are the average of each sensitivity and specificity calculated for each of the models created from the training sets.

Table 9: Cross Validation Results on the Logistic Regression Equation.

Model	Average Sensitivity	Average Specificity
Logistic Regression	0.7981	0.7940

The average sensitivity for was found to be 0.7981 while the average specificity was found to be 0.7940. This indicates that overall model of logistic regression can accurately be used to determine response by consistently producing good values of sensitivity and specificity. Each time a model was used in the cross validation, a new threshold was found to determine the responders from the non-responders. Table 10 shows the mean and the standard deviation for the threshold, sensitivity and specificity of the 2000 models used in the cross validation. The mean threshold value was found to be 0.4642 and had a standard deviation of 0.1678. This shows that during the 2000 models tested, the threshold value that was used remained relative the same in each of

the models since the standard deviation was small. The average sensitivity and average specificity, as previously mentioned, have shown that overall the model consistently was able to predict true responders and true non-responders wince the values of the averages were high. Also the standard deviation, for both the sensitivity and specificity were very small being on the order of thousandths. This shows that in the large number of models that were used, the specificity and sensitivity remained at the large value. This shows once again that the logistic regression model could be used to determine responders from non-responders.

Table 10: Average and Standard Deviation of Threshold, Sensitivity, and Specificity from Cross Validation.

Value	Average	Standard Deviation
Threshold	0.4642	0.1678
Sensitivity	0.7981	0.0096
Specificity	0.7940	0.0067

3.4 Results from Analysis of GTV 20% Threshold

Table 11: Results from Wilcoxon Rank-Sum Test (p-value) and AUC of the calculated ROC curve on SUV Metrics with >20% metric.

SUV Metric	p-value	AUC
Max	0.323	0.6364
Mean	0.109	0.7273
Highest 10%	0.048	0.7677
Highest 20%	0.048	0.7374
Highest 30%	0.068	0.7172
Highest 40%	0.068	0.7172
Highest 60%	0.081	0.7374
Highest 80%	0.095	0.7475
Highest 90%	0.149	0.7677
Combination of 10% and 20% Metrics	0.058	0.7575

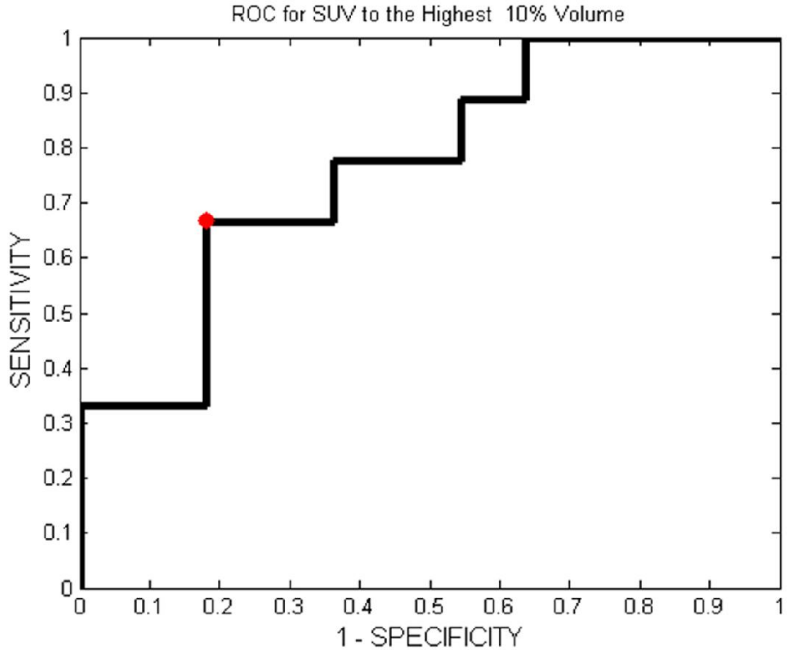
Table 11 shows the results from the Wilcoxon rank-sum test and ROC analysis. The most commonly used metrics of the maximum SUV and mean SUV were not found to be significant. The SUV to the highest 10% and 20% intensities were found to be significant with a p-value of 0.48. The SUV to the highest 10% had the larger AUC with a value of 0.7677 while the highest 20% AUC was 0.7373. Since the highest 10% had the higher value, it would be considered as the more promising indicator. In this case, the

combination of the two most significant metrics was not able to distinguish between responders and non-responders. The threshold values and their ability to distinguish responders (sensitivity) and non-responders (specificity) are shown in table 12.

Table 12: Threshold, Sensitivity, and Specificity for Significant SUV Metrics with >20% threshold.

Metric	Threshold	Sensitivity	Specificity
Highest 10%	0.4596	0.6667	0.8182
Highest 20%	0.5353	0.6667	0.8182

The threshold values for both the highest 10% and 20% were almost around a 0.5 fractional decrease from the pre-PET to the intra-PET while both metrics had the same sensitivity and specificity of 0.6667 and 0.8182 respectively. Figure 10 shows the ROC curves for the highest 10% and 20% metrics with the 20% threshold being used to create the GTV. The red marker indicated the cutoff point that distinguishes the best sensitivity and specificity that were described in table 12. Both of the curves in figure 10 have the desired bend to the upper left hand corner. The bend in the highest 10% is slightly more than that of the highest 20% which is expected since it has the larger AUC value.



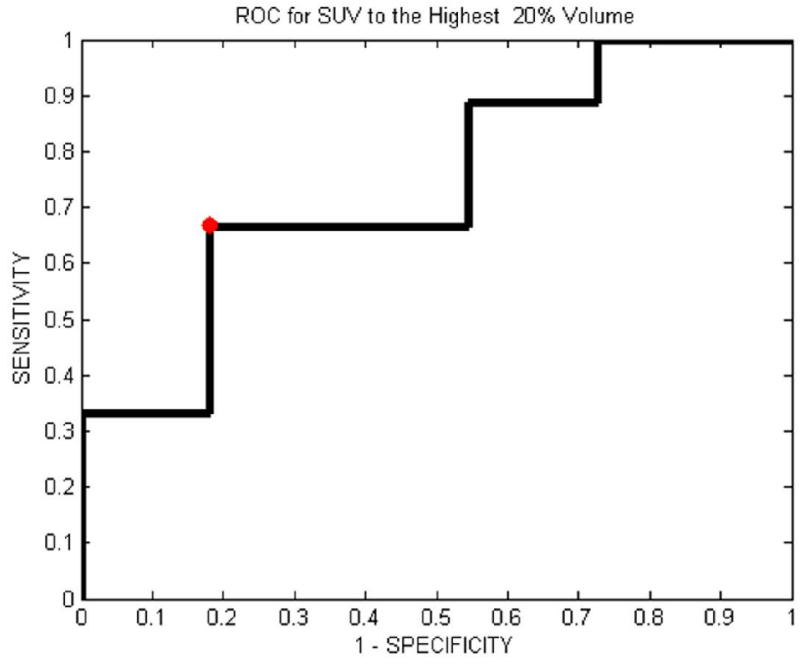


Figure 10: ROC Curves for Significant SUV Metrics Using a >20% Threshold

Table 13: Results from 3-Fold Cross Validation of the SUV Metrics Using >20% Threshold.

SUV METRIC	SENSITIVITY	SPECIFICITY
HIGHEST 10%	0.7778	0.7727
HIGHEST 20%	0.7778	0.5909

Table 13 shows the results from the 3-fold cross validation of the significant SUV metrics using the >20% threshold to contour the GTV. The SUV to the highest 10% metric was able to show that it could be used on unseen data and still be able to distinguish between responders and non-responders. The SUV to the highest 20% was able to distinguish between responders with a high specificity but was not able to distinguish between non-responders very well.

4. Discussion

From the results with >40% used as a threshold, it can be seen that the more commonly used SUV metrics of the maximum and mean SUV values were found to be not significant in distinguishing between responders and non-responders. Instead several other metrics were found to be better including the volumes of the highest 10%, 20%, and 30% intensities. These metrics were able to distinguish between responders and non-responders with a p=0.04 while the SUV maximum p-value was only 0.095 and the SUV mean had a p-value of 0.81 based on a Wilcoxon rank-sum test of the fractional differences in these metrics. While these p-values are small they do not meet the criteria of being significant ($p < 0.05$) and thus do not do a good job at telling apart the two distributions that were the responders and non-responders. The three significant SUV metrics all had a relatively large area under the ROC curve with a value of 0.7778 for each metric..

By using a combination metric that combined the three significant other significant SUV metrics, an even more significant indicator was able to be found. The simple combination of taking the average between the fractional decrease between the SUV to the highest 10% and 20% and also a combination from taking the average of the fractional decrease in the highest 20% and 30% produced results that were even more significant than three previously determined significant SUV metrics. Both of the combination metrics had a p-value of 0.033, lower than that of the 0.04 of the individual

metrics. By using cross validation on all of these significant metrics, the ability to reproduce an indication of response on unknown data was able to be determined. While all of the metrics were able to consistently produce a high level of sensitivity and specificity, the combination metric of the highest 10% and 20% metrics was able to provide the most robust result with a sensitivity of 0.8333 and a specificity of 0.7727.

While there were several SUV metrics that were able to determine the response of the patients to the treatment, only one texture feature was found to be significant. This feature was complexity and it had a p-value that was same as the individual SUV metrics ($p = 0.04$). Also the area under the curve for complexity was also the same as it was for the three individual significant SUV metrics with a value of 0.7778. Where this metric distinguishes itself from the individual SUV significant metrics is from its results from cross validation. This yielded a sensitivity of 0.8333 and a specificity of 0.7273 which were higher than that of the three individual SUV metrics which had a sensitivity of 0.7143 and specificity of 6400 but still smaller than that of the combination of the highest 10% and 20% metric.

The reason that the only texture feature was found to be complexity could be that the GTV contours that were used in this study did not offer a larger enough variation in the different intensities of the gray tones. Almost all of the computation of these texture features relies on there being a large difference in the intensities of a pixel and its neighboring pixels. Since the GTV contour was located in the tumor where there is a

large amount of uptake of the radiotracer, the intensity values of the gray tones would be very similar. The only feature that was found to be able to indicate response was the fractional decrease in complexity feature. The possible reason why this result was able to distinguish response where the others failed is that complexity relies on the visual information and the basic patterns of the image. If the basic pattern in the GTV changed enough during the course of treatment, this could indicate how well the patient is responding. More work could be done to use large contours that have a larger distribution of gray tone intensities to see whether other texture features based on the NGTDM would be able to be indicative of response.

The most statistically significant result from the Wilcoxon rank sum test was produced by doing the logistic regression analysis. This combined the data from the two separate analyses, SUV and texture features, to find the probability that the outcome will be a non-responder. Using these probabilities in comparison with the pathologic reports resulted in a p-value of 0.004 from the Wilcoxon rank sum test, well below the 0.05 significance criteria. This p-value was much lower than that of the other values tested using the Wilcoxon rank sum test by a magnitude of 10. While some of the other values were close to 0.05 level the logistic regression p-value was nowhere close to being near it. The area under the ROC curve for the logistic regression analysis also proved to be the best of the significant values with an AUC of 0.8889. This means that this equation had almost an 89% probability of correctly identifying the response to the treatment.

Cross validation of this model also proved to be substantial. Since there was a total of 84 different combinations of breaking the 9 responders into groups of 6 and 330 combinations of breaking the 11 non responders into groups of 7, there would be a possibility of over 27,000 different training and test models that would have been able to be used. Since this that many models would have been a bit of an "overkill", only 2000 models were used to for the validation. This was still a substantial amount of models to be able to test the how the logistic regression model would react to unknown data. By taking the average sensitivity and specificity collected from each of the 2000 test sets, the robustness of this model was brought to light. The average sensitivity was 0.7981 and the average specificity 0.7980. Both of these values show that the model when shown unseen data, was still able to accurately predict the correct response to the treatment. Another good statistic that was produced from the random splitting cross validation was the standard deviation of the sensitivity and specificity. Both of these values were on the order of a thousandth showing little deviation between each test. This shows that all of the models produced a sensitivity and specificity that was around 0.79. The threshold for the probability also did not deviate much as well but it did more so than the sensitivity and specificity. The average threshold that was used to test each model was higher than the threshold that was found in the original logistic model.

Overall it appears that when the significant metrics were combined together in some form, they were able to better distinguish response. In the SUV metrics, both

combination terms used resulted in lower p-values and higher area under the ROC curves than the three individual metrics. Also, when both the SUV analysis and the texture feature analysis were combined using logistic regression, an even more significant result was produced. This combination resulted in the lowest p-value and highest AUC than any of the other metrics tests.

For when >20% of the maximum SUV threshold was used to contour, only two metrics were found to show promise in distinguishing response. These values were the highest 10% and 20% metrics. While both of these values had the same resulting p-value, the highest 10% metric had a larger AUC. This hinted that it may be the better value to be used if 20% is used as a threshold. Further evidence that this was the better metric came from the cross validation of these two metrics. The 10% metric was able to distinguish unseen data with a high sensitivity and specificity while the 20%'s specificity was just over the 50% random guess. From these results, SUV to the highest 10% would be the best SUV metric. Once again the more commonly used maximum and mean SUV metrics were not found to be able to distinguish a response.

Comparing the two different threshold levels used for creating the PET based GTV. It appears that the 40% threshold was better at indicating pathological response. It had several different SUV metrics that were much more significant than that of the 20% threshold that had metrics that were just significant. Also when comparing the two thresholds it can be seen that there were similar significant metrics. These metrics were

the SUV to the highest 10% and 20%. Since these values were significant over multiple thresholds, there could be a possibility for them to significant across a multitude of threshold values. If a study were to come out that would find the best threshold to be used to delineate a PET based GTV, SUV to the highest 10% and 20% might be able to determine response.

5. Conclusion

The goal of this study was to utilize F-18 FDG PET/CT scans to determine the response in esophageal cancer patients during radiation therapy treatment. The fractional decrease in several SUV metrics for a >40% threshold were able to distinguish responders from non-responders when compared to the post-treatment pathological reports. These metrics were the SUV to the highest intensity 10%, 20%, and 30% volumes. Although each one was significant, it was not possible to determine which one of these metrics was the better individually at determining response. When the individual metrics were combined together, a more significant result was found. The combination of the highest 10% and 20% volumes and the combination of the highest 20% and 30% volumes produced the lowest p-values and highest area under the ROC curves. Cross validation not only revealed the robustness of these results but also revealed which of these metrics would be better overall at determining the response to treatment. The combination of the highest 10% and 20% volumes appeared to be the best SUV metric in determining response since it not only had the lowest p-value and highest AUC, but also had the best sensitivity and specificity results from the cross validation. This metric used a threshold of 0.4319 to distinguish responders from non-responders with a sensitivity of 0.7778 and a specificity of 0.7273. While all of these metrics proved to be indicative of response, the more commonly used SUV metrics (maximum and mean SUV) were found to be insignificant at determining response. Further work could

be done to investigate if any other combination of significant SUV metrics would yield an indicator. Also when SUV 20% was used to contour the PET based GTV, SUV to the highest 10% and 20% were significant. Cross validation proved that of these two values the 10% metric showed more promise in being an indicator for response. Since this metric was significant in both threshold cases, there is a possibility that this metric could be used to determine response across a range of thresholds.

Overall, the best results from the study were produced when the most significant SUV metric and the complexity texture feature were both combined to predict a response. Using logistic regression analysis an equation was created that both the combination metric of the 10% and 20% volumes and the texture feature complexity could be used as inputs to accurately predict response. Not only did this equation accurately predict the response to the treatment, but random splitting cross validation was able to be utilized to show that this model would consistently produce a high level of sensitivity and specificity. From a Wilcoxon rank-sum test on the predicted probability values the optimal threshold for predicting responders from non-responders was 0.3186. Anything probability that was greater than this value would be determined as a non-responder while any value less than it would be deemed a response. Ideally when using a logistic regression for a prediction, a probability of 0.50 is used to determine the split of the prediction value. A possible reason why a lower value was found to be best in this study could be that there are so often local failures in the

treatment of esophageal patients. Another reason why this value is low is that this study was limited to a small sample of patients. More patients could result in an optimal threshold value that would be closer to the 0.50 value. This could be seen from the cross validation results on the logistic regression model as the average threshold from the 2000 test models was 0.4642. Future work could be done to test if other predictor models or other input metrics could be used to accurately predict the response to treatment. Also, further analysis on the logistic regression model using the complexity and the combination metric of the highest 10% and 20% as inputs could be done with a larger patient study.

References

1. Zhang Y. Epidemiology of esophageal cancer. World J Gastroenterol. 2013;19:5598–5606.

2. Howlader N, Noone AM, Krapcho M, Garshell J, Neyman N, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2011, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2011/, based on November 2013 SEER data submission, posted to the SEER web site, April 2014.

3. Parkin, D. Max, et al. "Global cancer statistics, 2002." *CA: a cancer journal for clinicians* 55.2 (2005): 74-108.

4. Welsh, James, et al. "Failure patterns in patients with esophageal cancer treated with definitive chemoradiation." *Cancer* 118.10 (2012): 2632-2640.

5. Downey, Robert J., et al. "Whole body 18FDG-PET and the response of esophageal cancer to induction therapy: results of a prospective trial." *Journal of Clinical Oncology* 21.3 (2003): 428-432.

6. Roedl, Johannes B., et al. "Adenocarcinomas of the esophagus: response to chemoradiotherapy is associated with decrease of metabolic tumor volume as measured on PET–CT: comparison to histopathologic and clinical response evaluation." *Radiotherapy and Oncology* 89.3 (2008): 278-286.

7. Swisher, Stephen G., et al. "2-Fluoro-2-deoxy-D-glucose positron emission tomography imaging is predictive of pathologic response and survival after preoperative chemoradiation in patients with esophageal carcinoma." *Cancer*101.8 (2004): 1776-1785.

8. Zhong, Xiaojun, et al. "Using 18 F-fluorodeoxyglucose positron emission tomography to estimate the length of gross tumor in patients with squamous cell carcinoma of the esophagus." International Journal of Radiation Oncology* Biology* Physics 73.1 (2009): 136-141.

9. Biehl, Kenneth J., et al. "18F-FDG PET Definition of Gross Tumor Volume for Radiotherapy of Non–Small Cell Lung Cancer: Is a Single Standardized Uptake Value Threshold Approach Appropriate?." Journal of Nuclear Medicine 47.11 (2006): 1808-1812.

10. Lucignani, G., G. Paganelli, and E. Bombardieri. "The use of standardized uptake values for assessing FDG uptake with PET in oncology: a clinical perspective." Nuclear medicine communications 25.7 (2004): 651-656.

11. Kadoya, Noriyuki, et al. "Evaluation of various deformable image registration algorithms for thoracic images." Journal of radiation research (2013): rrt093.

12. Amadasun, Moses, and Robert King. "Textural features corresponding to textural properties." Systems, Man and Cybernetics, IEEE Transactions on 19.5 (1989): 1264-1274.

13. Hollander M, Wolfe DA. "Nonparametric Statistical Methods". New York, NY: John Wiley & Sons; (1973)

14. Wilcoxon, Frank, S. K. Katti, and Roberta A. Wilcox. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. American Cyanamid Comp., 1963.

15. Zweig, Mark H., and Gregory Campbell. "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." Clinical chemistry 39.4 (1993): 561-577.

16. Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology 143.1 (1982): 29-36.

17. Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

18. Fernando, S., et al. "Using FDG-PET to delineate gross tumor and internal target volumes." International Journal of Radiation Oncology* Biology* Physics 63 (2005): S400-S401.