THE UNIVERSITY OF CHICAGO

REDEFINING RADIOLOGIC RESPONSE FOR PATIENTS WITH MALIGNANT PLEURAL MESOTHELIOMA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MEDICAL PHYSICS

BY
ZACARIAH EVAN LABBY

CHICAGO, ILLINOIS
JUNE 2012

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

To my family, whose decency and support have made me all that I am.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It takes many ants to create an anthill, though only one ant can stand at the very top. So it is with this dissertation, and I hope to give some meager thanks to some of the people who have helped me build my own anthill of knowledge.

My first and most important acknowledgment is to my thesis adviser, Samuel G. Armato III, Ph.D. Sam's inquisitive and helpful guidance has given me both the direction and berth required to mature as an independent investigator, for which I am most grateful. Next, let me acknowledge the other members of my dissertation committee: Maryellen Giger, Ph.D.; Hedy Kindler, M.D.; Christopher Straus, M.D.; and James Dignam, Ph.D. Their analytical, clinical, radiological, and statistical guidance (respectively) have been instrumental in the completion of this work, which spans multiple fields. I would also like to specifically thank Charles Metz, Ph.D., for his enlightening conversations and shared wisdom throughout my time at the University of Chicago.

I'd like to thank Anna Nowak, M.D., Ph.D., and the rest of the kind people at Sir Charles Gairdner Hospital in greater Perth, Australia. Her assistance in creating the main cohort of patients used herein is key to this work, and she and the rest of the people with whom I interacted in February of 2011 made my work there easy and fruitful. Getting closer to home, I'd like to thank the members of my lab, William Sensakovic, Ph.D., Adam Starkey, Alexandra Cunliffe, and Neal Corson, Ph.D., for their help and conversations. Lorenzo Pesce, Ph.D., played an important role in my understanding of applied statistics, and many other friends in my department have contributed to my experience in Chicago, including (but not limited to) Elizabeth Hipp, Erik Pearson, Phillip Vargas, and Anita Dhyani. These individuals have influenced not only my academic experience, but have been teammates and allies on what was sometimes a difficult journey.

I'd also like to thank the faculty in the Radiation Oncology department, especially Chester Reft, Ph.D. and Bulent Aydogan, Ph.D. With their help, I have been able to keep one foot in the door of the clinical physics realm. They have helped me define my future career goals, and I appreciate the opportunities they have given me.

Finally, I'd like to thank my family. Throughout my life, the love and support of my mother Kristine, my father Daniel, and my sister Kira have been the forces that have shaped me into what I am today. My sister and I would not have been able to succeed as we have in academic pursuits without the foundation of intelligence and hard work given to us by our parents. My wife Kristin has been my closest friend, ally, and partner, and if this work exemplifies half the skill she has as a scientist, I would be pleased. I look forward to our first co-authored manuscript.

I know this achievement, this anthill, reflects only the uncountable grains of sand carried by others, and my most sincere hope is that they understand my gratitude.

ABSTRACT

The current standard for medical image-based tumor response assessment for patients with malignant pleural mesothelioma is the modified Response Evaluation Criteria In Solid Tumors (RECIST) measurement technique with changes classified according to the standard RECIST response classification criteria. While the modified RECIST measurement technique was developed specifically for the unique morphology and growth patterns of mesothelioma, the standard RECIST classification criteria are used across a wide range of diseases and are certainly not specific to any one disease. As a coherent progression, this work extends the definition of radiologic response for patients with malignant pleural mesothelioma from one to three dimensions, both in the context of a discrete classification system and as a continuous prognostic model. Furthermore, temporal and spatial measurements are combined in an investigation of dynamic imaging for patients with mesothelioma.

This work begins by quantifying the association between standard tumor response classification from thoracic computed tomography (CT) scans and patient survival in a study focused on the standard linear thickness measurement technique. The association between tumor response and patient survival was improved by identifying new response classification criteria specifically for a database of treated mesothelioma patients. In place of summed linear thickness measurements, summed measurements of disease area from a limited number of CT sections were investigated as a more complete and potentially less variable metric for tumor response assessment. However, the inter-observer variability in such measurements, even in a constrained follow-up contouring task, was too broad for reliable use as a response assessment technique in patients with mesothelioma. Next, full volumetric segmentations of pleural disease volume were investigated as a response assessment measurement technique. A comprehensive model to predict patient survival was built using time-changing measurements of disease volume in conjunction with other clinical covariates, where it was shown that increasing pleural disease volumes are significantly associated with poor patient prognosis in both univariate and multivariate prognostic models. Automatically segmented

lung volumes were also investigated as an alternate response assessment measurement technique. Because of the anatomy associated with mesothelioma, when the disease volume increases, it is reasonable to believe that the aerated lung volume will decrease correspondingly. Lung volume segmentation is a completely automated process and is a computationally simpler task than pleural disease volume segmentation. As expected, decreases in lung volume were shown to be significantly associated with poor patient prognosis, and the performance of the prognostic model using lung volume measurements was statistically nearly identical to the model using disease volume measurements. Both models were compared with a model using summed linear thickness measurements, and the performance of the linear measurement model was on average better than the performance of either the disease volume or lung volume models (though not significantly so). This work provides the first evidence that trajectories of three distinct response assessment measurement techniques are significantly associated with mesothelioma patient survival in univariate and multivariate models. Finally, this work reports the results of a pilot study on dynamic contrast-enhanced (DCE) CT for patients with mesothelioma. A DCE-CT imaging protocol was developed to dovetail with a clinically indicated standard chest CT scan, and software tools were developed to analyze DCE-CT parameters from volumetric regions of interest in a spatially co-registered reference frame. DCE-CT parameter values are reported for individual scans and changes in DCE-CT parameters are calculated for two cohorts of patients, one undergoing treatment and the other on observation. Only 13 patients have received both DCE-CT scans; the initial pilot study will continue until a final accrual goal of 20 patients with two DCE-CT scans each. This dissertation investigated a variety of new response assessment strategies for patients with malignant pleural mesothelioma. These techniques will hopefully impact the tools clinicians use to assess patient response in both phase II clinical trials and routine patient care.

CHAPTER 1

INTRODUCTION

> "In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be." – Sir William Thomson (Lord Kelvin)

1.1 Mesothelioma

Malignant pleural mesothelioma (MPM) is a malignancy of the pleural membranes which surround the lungs and separate them from the thoracic wall and is primarily caused by exposure to asbestos [1,2]. Mesothelioma in general is not exclusive to the thoracic cavity but can appear in the body where mesothelium cells are present, which often make up all or part of the protective linings surrounding internal organs and separating body cavities. For instance, peritoneal mesothelioma can occur in the lining of the abdominal cavity, though approximately 90% of malignant mesotheliomas are of pleural origin [3]. For the purpose of this work, "mesothelioma" (or MPM) will be taken to mean malignant pleural mesothelioma.

1.1.1 Asbestos

The relationship between exposure to asbestos and development of mesothelioma is well-documented [1, 3]. Asbestos is a mineral with properties that work well as an insulating construction material, and asbestos has been used in wall and pipe insulation, floor tiles, ceilings, walls, con-

crete fences, and other places. The asbestos fibers are long and thin crystals, and these crystals can be inhaled during uncontrolled exposure to loose airborne fibers. The exact mechanism by which asbestos is carcinogenic is not fully understood, but it is believed that the airborne asbestos fibers become lodged in the alveoli, and larger fibers migrate through the lung parenchyma to the pleural lining. There, the fibers repeatedly scratch and irritate the pleural surfaces with every breathing cycle, and the constant damage repair processes of the immune system are thought to result in DNA strand breaks, eventually leading to mutation and carcinogenesis. Different types of asbestos have different crystalline shapes: chrysotile, or white, asbestos is described as "curly," whereas amosite (brown) or crocidolite (blue) asbestos have more needle-like crystals. Because of the mechanisms believed to be responsible for mesothelioma carcinogenesis, amosite and crocidolite asbestos are jointly believed to be responsible for the majority of disease incidence [4]. A photograph of asbestos fibers used to weave insulating material is shown in Figure 1.1.

Because asbestos exposure is usually occupational, and males were largely involved in the mining and construction uses of asbestos, the incidence of mesothelioma is higher in males than in females. The annual incidence is estimated in the range of 15–20 cases per million, with less than five cases per million occurring in females (contrasted with an annual incidence of approximately 560 per million for breast cancer or 570 per million for prostate cancer) [5, 6]. The latent period of the disease is around 30–40 years, and asbestos use reached its peak in the US in the late 1960s, explaining the relatively high incidence of the disease now compared with past decades [3, 7, 8]. While evidence suggests that the disease may have peaked in the US in the late 1990s or early 2000s, European incidence is not forecast to peak until sometime this decade, and incidence in countries that continue to use asbestos in new buildings will continue to increase. For instance, asbestos use continued longer in Europe, Japan, and Australia, and incidence is expected to peak in these regions in 2020, 2025, and 2015, respectively. Therefore, while the US incidence of mesothelioma is relatively low, the disease is not "going away" anytime soon. The Australian incidence is one of the higher incidences worldwide at approximately 40 per million, largely due to

Figure 1.1: A photograph of asbestos fibers used to weave insulation material. Note the long, thin crystal shape. Source: Wikimedia, "Asbestos fibres," accessed January 31, 2012 (http://commons.wikimedia.org/wiki/File:Asbestos_fibres.jpg; Wikimedia Commons is a media file repository making available public domain and freely-licensed educational media content to everyone).

Figure 1.2: A sample of blue (crocidolite) asbestos from Wittenoom in Western Australia. Note the fibrous and spindle-like nature of the mineral. The ruler is 1 centimeter. Source: Wikimedia, "Blue asbestos (teased)," accessed January 31, 2012 (http://commons.wikimedia.org/wiki/File:Blue_asbestos_(teased).jpg).

the presence of vast mineral mines in Australia, especially mines like Wittenoom Gorge in Western Australia, where asbestos tailings were used as covering spread on playgrounds and schoolyards (in addition to the airborne asbestos from the mining operations) [1]. A sample of blue asbestos from Wittenoom is shown in Figure 1.2.

1.1.2 Treatment

For patients with resectable MPM, the primary means of treatment is surgery. Surgical interventions include thoracoscopy (resection of masses within the pleural cavity, often video-assisted), pleurectomy/decortication (P/D: open thoracotomy with stripping of pleural lining from hemithoracic and mediastinal structures), and extrapleural pneumonectomy (EPP: removal of entire tissues from the affected hemithorax, including pleura, lung, lymph nodes, etc) [9]. Because EPP is a

more radical treatment option, its use is somewhat controversial; the procedure is associated with significant morbidity, and it has not been shown to improve long-term survival [10]. Surgical intervention is still associated with recurrence, both distant and local. For P/D, local recurrence dominates, with rates from 64% to 72%, and for EPP, distant recurrence dominates, with rates from 41% to 44% [11–13].

Radiation therapy also plays an important role in the treatment of MPM. The primary use of radiotherapy in treating MPM is for prophylaxis, either at the site of surgical incision to prevent disease spreading, or to the entire hemithorax after EPP. Both approaches aim to kill microscopic disease left behind by surgical intervention. Definitive treatment of the entire hemithorax after EPP involving a radiation dose of 54 Gy improved locoregional control over moderate radiation doses of 30 Gy, but the difference between dose cohorts was not statistically significant. Furthermore, intensity modulated radiation therapy (IMRT) resulted in a local control rate of 87% but had associated high toxicity, including a case of fatal radiation pneumonitis [14, 15].

Chemotherapy can be administered in the neoadjuvant or adjuvant settings for resectable MPM. The most common drugs used are a platinum-based agent (cisplatin or carboplatin) with either gemcitabine or pemetrexed. A Swiss multi-center trial of neoadjuvant cisplatin plus gemcitabine prior to EPP showed an overall median survival of 19.8 months for all patients, with median survival for patients with resectable disease (74% of patients) increasing to 23 months [16]. For patients without resectable disease, chemotherapy is used as a front-line treatment. The current standard is cisplatin in combination with pemetrexed, which became the standard after a large trial ending in 2003 [17]. This regimen showed an overall median survival of 12.1 months. Cisplatin has also been combined with gemcitabine for overall median survival times from 9.4 to 13 months (see Table 2 of reference [9]). Carboplatin is sometimes substituted for cisplatin, since carboplatin reduces the side-effects of nausea, vomiting, and kidney toxicity compared with cisplatin. However, carboplatin is not as potent as cisplatin [18]. Both agents use platinum to interfere with DNA mechanisms, including mitotic cell division, and when repair is not possible, the cells default to

apoptosis, or cell death.

1.1.3 Imaging

The standard diagnostic imaging modality used in the assessment of mesothelioma is computed tomography (CT) [19–23]. While patients may be initially evaluated at presentation with a projection chest x-ray (CXR), CT is the primary modality for diagnosis, staging, and preoperative evaluation. These CT scans are typically performed with an iodinated (radio-opaque) contrast agent unless the contrast agent is not tolerated by the patient (for instance because of renal insufficiency). Whether or not contrast media is used, the presentation of MPM on CT scans can pose challenges. The Hounsfield Unit (HU), sometimes called the CT number, values of mesothelioma overlaps considerably with the adjacent chest wall musculature [24]. The overlap in HU values can be seen in Figure 1.3, where the signal of the disease appears very similar to the adjacent muscle. Furthermore, pleural effusion (PE) is also commonly observed with mesothelioma, and the HU values of pleural effusion can appear very similar to disease tissue, depending on the fluid content (protein concentration, etc). Especially when very small pockets of PE are mixed with solid tissue, the partial volume effect of CT imaging can make distinguishing PE from disease virtually impossible.

MRI can be used for MPM imaging and has the advantage of improved soft-tissue contrast over CT, depending on the pulse sequence (or weighting) of the scan [23]. This additional image contrast leads to better identification of chest wall invasion, diaphragmatic invasion, and mediastinal invasion. However, MRI imaging is more susceptible to imaging artifacts because of the extended duration of the scans, with aliasing especially common.

Fluorodeoxyglucose positron emission tomography (FDG-PET) also provides complementary imaging information to CT scans. MPM is FDG-avid, meaning that because of the metabolic properties of the tumor, FDG shows preferential uptake in the diseased tissue [25–28]. This property, along with the known FDG avidity of affected lymph nodes and metastatic masses, has great value in the staging process for MPM. Complete image staging, including primary tumor staging, node,

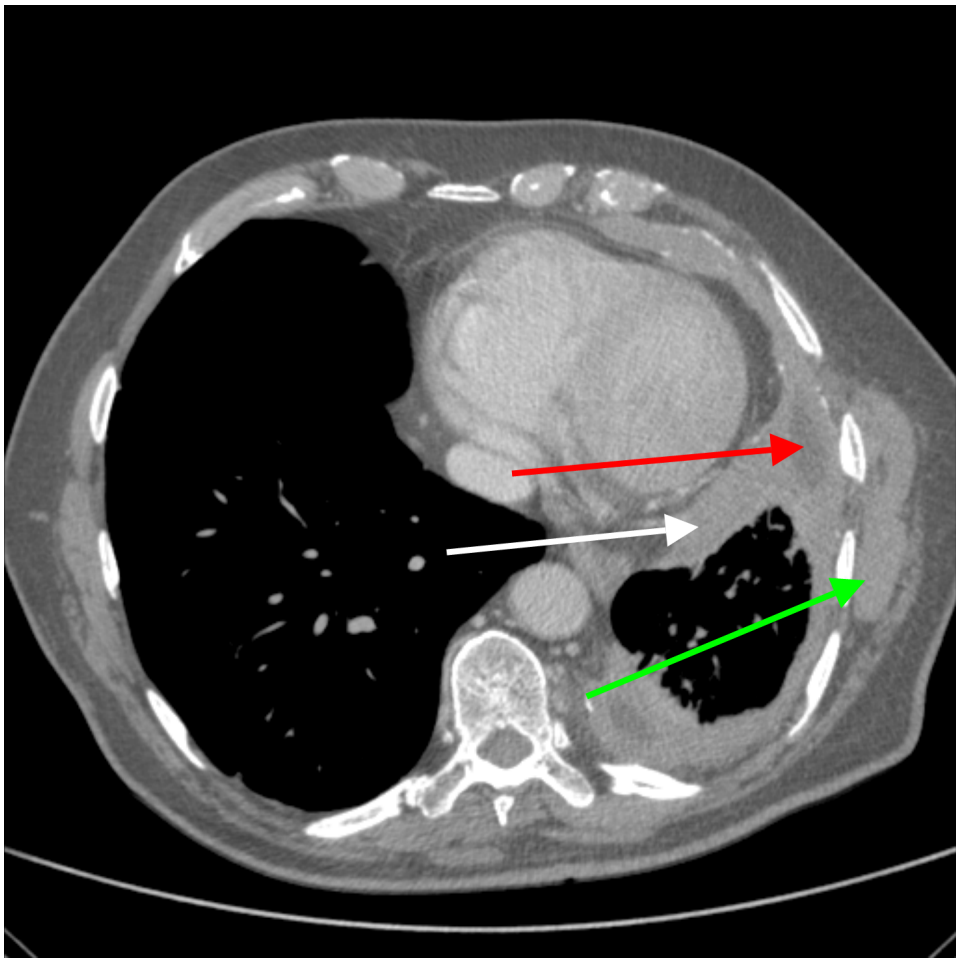


Figure 1.3: Typical presentation of mesothelioma on CT. The white arrow indicates disease, the red arrow indicates pleural effusion, and the green arrow indicates adjacent musculature. Note the decidedly aspherical disease morphology.

and metastasis staging (TNM), is often impossible without an FDG-PET scan [29]. However, the spatial resolution of PET imaging (typically on the order of 1 cm) is not comparable with the spatial resolution of MRI or CT (typically on the order of 1 mm), and even the CT used for attenuation correction in PET imaging is typically not of diagnostic quality. However, there has been interesting work on the use of FDG-PET metrics for tumor response in MPM, including the standardized uptake value (SUV), which is a normalized quantitative metric for FDG uptake [28].

1.1.4 Prognostic Markers

Prognostic markers are pieces of information that have value in forecasting the likely future outcome for a patient. In a very broad sense, prognostic markers are useful in making predictions about a patient's survival, whether the markers are taken before or during the course of a patient's treatment. There is some ambiguity between the terms "prognostic" and "predictive" in the realm of patient outcomes, with "predictive" usually implying a marker that has value in predicting how well a patient will respond to a specific therapy, and "prognostic" implying a marker that has value in forecasting future survival in some way, sometimes from baseline marker values only. In this work, the term "predictive" will generally be avoided to describe patient-specific markers, since in a broad sense the term "prognostic" encompasses the definition of "predictive."

In MPM, several studies have focused on the identification of prognostic markers [30–35]. Some common features of these studies are patient age, sex, performance status (PS) as measured on the Eastern Cooperative Oncology Group (ECOG)/World Health Organization (WHO) scale, white blood cell (WBC) count, blood platelet count, histological subtype, presence of chest pain, and weight loss. Specifically, groups with worse prognosis are older patients, males, ECOG PS 1 and 2 (compared with 0), high WBC count ($> 8.3 \cdot 10^9$ per liter), high platelets ($> 400 \cdot 10^9$ per liter), sarcomatoid or biphasic histology (compared with epithelioid), confirmed chest pain, and confirmed weight loss. These factors (and various others) have been used to create discrete prognostic groups, such as the six groups of the Cancer and Leukemia Group B (CALGB) system,

or the two groups in the European Organization for Research and Treatment of Cancer (EORTC) system [31, 32]. Both of these models, as well as the individual prognostic markers, have been validated [33]. Using the same EORTC patient cohort, a separate model has been developed using histology, disease stage (stage IV compared with I, II, or III), and ECOG PS [34]. There has also been recent interest in blood serum markers such as soluble mesothelin concentration [35].

Some image-based features have proven to be prognostic for MPM as well. In a 1998 study, Pass *et al.* [36] showed that preoperative tumor volume, as measured using manual contours on CT scans, was prognostic for survival. Median survival for patients with preoperative volumes less than 100 cc was 22 months, while for patients with preoperative volumes greater than 100 cc the median survival was 11 months ($p = 0.03$). FDG-PET imaging has also been used as a prognostic marker [25, 28, 35, 37, 38]. The general trend of these models is that larger values of SUV (or occasionally maximum SUV) and/or total glycolytic volume (TGV), a measure combining the metabolic activity and volume of the entire tumor mass, are associated with poor patient prognosis.

1.2 Measuring Disease Response to Treatment

In general, the concept of an "effective treatment" hinges on the definition of "effective." The ultimate standard for efficacy is patient survival; effective treatments should by definition prolong patient survival. However, because of the expected survival for some diseases, measuring efficacy through survival can be prohibitively time-consuming. Disease response metrics are common surrogates for patient survival in the assessment of treatment efficacy, and as such, response metrics are vital components of patient care and clinical trials.

Response metrics measure some aspect of the patient's disease, often (but not always) a measure of tumor burden. These measurements are then used as a more immediate tool to assess the efficacy of therapy. Is the patient's disease shrinking? If so, they may be responding to the treatment. Is the disease burden increasing? The patient's disease may be progressing, potentially indicating a lack of therapeutic efficacy. As Nowak states in a 2005 article, "a decrease in tumor

size may or may not achieve palliation in individual patients. However, tumor response is an important surrogate for patient benefit in non-randomized clinical trials where symptom improvement and increased survival are difficult to assess" [39].

This type of "immediate" feedback is especially useful in phase II clinical trials, where the goal is simply to show that the treatment has demonstrable efficacy. The purview of a phase III trial is a definitive assessment of the novel therapy in comparison with the "gold standard." Patient survival remains the preferred measure of efficacy in phase III trials, but a phase II trial is greatly expedited through the use of surrogate markers such as disease response metrics. However, for the success of a phase II trial to have relevancy, the disease response metric must be a useful surrogate for patient benefit or survival. Recent analysis of clinical trials for chemotherapeutic agents showed a disappointing tendency for pharmaceuticals passing phase II trials to fail in a phase III trial, and perhaps part of the issue is an inappropriate system for response assessment [40, 41].

1.2.1 Historical Development of Response Models

Any image-based response evaluation method has two components; the first describes a protocol for making measurements, and the second describes how to classify patients into response categories once those measurements are available. The first widely utilized radiologic response assessment tool was the World Health Organization (WHO) bi-dimensional measurement technique [42]. Using this technique, the medical image section containing the largest tumor diameter was found, and linear measurements were taken of both the largest tumor diameter and the largest span perpendicular to the first measurement. The product of these two linear measurements was used as the quantity to represent tumor burden. If the product of the bi-dimensional measurement increased by more than 25% from the minimum of previous measurements (nadir), the patient was classified as having progressive disease (PD), and if the measurement decreased by 50% or more from the baseline measurement, the patient was classified as partially responsive (PR). Complete response was the disappearance of measurable disease, and stable disease (SD) was defined for measure-

ment changes lying between the previously mentioned classification groups. These -50%/+25% cutoffs for PR and PD, respectively, were based on previous breast cancer cohort studies, where the origin of the criteria was, in part, from what physicians believed to be reasonable palpable changes in tumor burden (i.e., the amount of change the physicians could reliably identify from palpation alone).

Later, the Response Evaluation Criteria In Solid Tumors, or RECIST criteria, were derived to simplify this measurement process; the two necessary measurements were reduced to one (again, the longest tumor diameter), and the cutoff criteria were derived from a geometrical relationship [43]. For a sphere where the cross-sectional area increases by 25%, the diameter increases by 12%, whereas a cross-sectional area reduction of 50% for the same sphere would correspond to a diameter decrease of 29%. The -29%/+12% "sphere model" extrapolations were rounded to -30%/+20%, leading to the current RECIST classification criteria [44, 45]. The RECIST measurement technique and classification criteria are currently used across many disease types including lung cancer, mesothelioma, breast cancer, colorectal cancer, prostate cancer, gastro-intestinal stromal tumors, soft tissue sarcomas, brain tumors, renal cell carcinomas, and others [46].

1.2.2 Response Assessment in Mesothelioma

The wording of the original RECIST measurement technique poses a serious challenge for MPM. Since MPM grows as a pleural rind surrounding the lung, does the "longest diameter" refer to the diameter of the entire lung? To the maximum chord across the annular ring? To the maximum thickness of the rind? The measurement technique for MPM was ill-defined until 2004, when the modified RECIST criteria were published [46–48]. With the modified RECIST technique, two linear "short axis" measurements, or thickness measurements, are summed from each of three axial sections. The sum of the six thickness measurements is tracked across time, and patient response classification is performed according to the same -30%/+20% criteria that are applied to many other diseases.

Whereas the modified RECIST measurement technique is better suited to MPM morphology, the classification criteria are still based on arbitrary standards with historical baggage. There is room for doubt on the applicability of such criteria for classification of response in a disease so typically aspherical as mesothelioma (see Figure 1.3) [49]. As the eccentricity of a tumor increases, the concordance between the RECIST and WHO systems has been shown to significantly decrease [50]. Indeed, others have derived theoretical response classification criteria for mesothelioma based on geometric models other than spheres, such as the lens, crescent, or annulus [51]. On the whole, these geometric models indicate that uni-dimensional measurements made according to the modified RECIST guidelines would be better classified with criteria where the definition of stable disease was more broad (for instance, in the crescent model, criteria of -66%/+74% would replace the current -30%/+20% standards). Broadening the definition of stable disease would also have benefit in the context of inter-observer variability: it has been previously shown that modified RECIST measurements have a relative inter-observer variability that can span a range of 30% under highly idealized image measurement conditions [52, 53]. With the standard RECIST -30%/+20% criteria, a patient with truly stable disease (i.e., 0% change) may be incorrectly classified as PR or PD due to observer effects alone.

While the goal of the (modified) RECIST system is to capture changes in overall tumor bulk, the true measure of tumor burden is three-dimensional volume. Numerous studies have investigated the use of tumor volume and tumor volume-related measurements for response assessment in patients with MPM using MRI, CT, and FDG-PET imaging [35, 37, 54–58]. The main challenge in these volumetric studies is the segmentation of the complete tumor volume. Using FDG-PET imaging, the segmentation of MPM is greatly facilitated by the FDG avidity of the tumor. Three-dimensional semi-automated region growing techniques can be used with high reproducibility to capture the complete tumor volume [55, 59], and even more rudimentary simple thresholding techniques also appear to perform well [56].

The segmentation of MPM in CT images is more problematic, and both published studies use a

semi-automated tool for MPM volume segmentation [57,58]. The study by Frauenfelder *et al.* [57] used a linear shape-based interpolation technique, requiring contours on "every fourth or fifth slice." The main conclusion of this study related to volumetric response was that the inter-observer agreement of volumetric response classification is much higher than for manual modified RECIST response classification (general $\kappa = 0.9$ vs general $\kappa = 0.33$, respectively) [57]. This may be a by-product of the fairly wide definitions given to the response categories for volumetric data. The study by Liu *et al.* [58] utilized a combination of semi-automated techniques for volumetric MPM segmentation, and their analysis revealed changes in tumor volume to be significantly associated with patient survival. Patients experiencing tumor growth had a median survival of 11.5 months, while patients with tumor decrease had a median survival of 18.1 months.

There has been progress toward an automated tool for the volumetric segmentation of MPM. The method developed by Sensakovic *et al.* [60] uses automated segmentations of the lung and hemithoracic boundaries, and the results agree with human observers to the extent that human observers agree with each other. The area overlap metric (AOM), defined as the volumetric overlap between two sets of contours divided by the total volume encompassed by *either* of the sets of contours, was used as the validation metric. The median AOM value amongst observers was 0.52, while the median AOM value between the computerized method and the manual observers was 0.48. While this places the automated method on par with the manual observers, the fact remains that neither AOM value instills much confidence.

With the challenges in volumetric MPM segmentation on CT images, it is not unreasonable to wonder why FDG-PET response assessment does not have a wider usage. Figure 1.4 highlights the difficulty of MPM segmentation on CT and the ease of MPM segmentation on FDG-PET. However, PET imaging is more expensive, has a lower spatial resolution, and is less accessible than CT imaging.

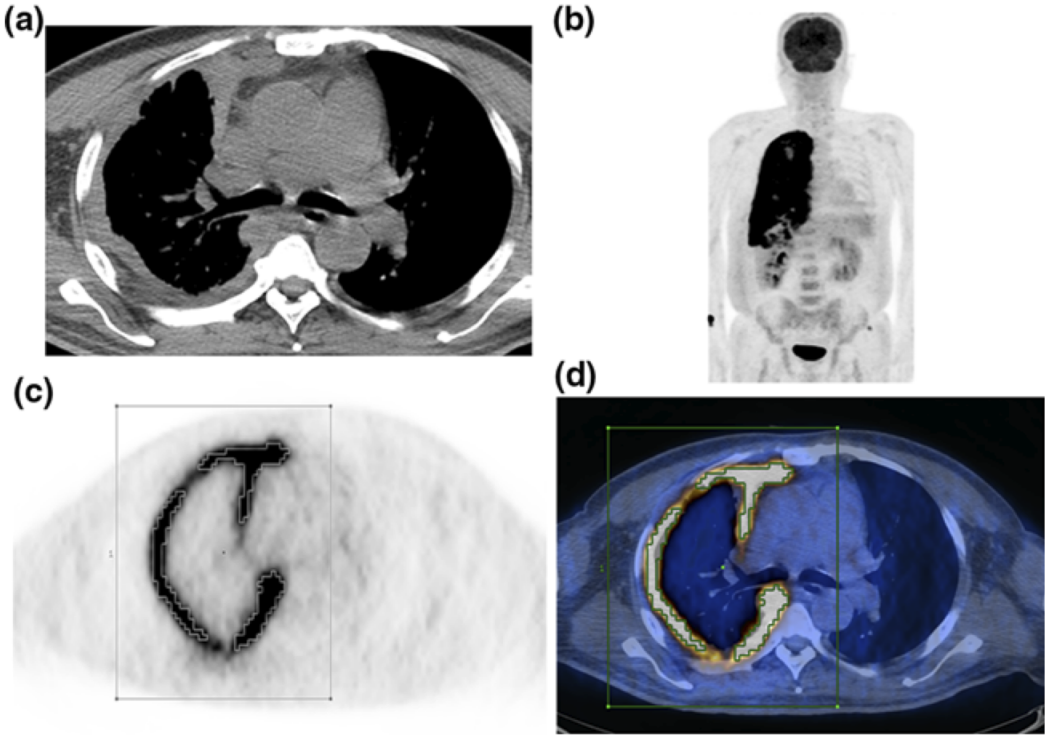


Figure 1.4: Appearance of MPM on axial CT (a) and FDG-PET (b,c). The coronal PET image (b) and the axial PET image (c) highlight the FDG-avidity of MPM. Automatic volume segmentation contour is shown in (c) and in a PET/CT fusion image (d). Taken from Figure 2 of reference [37] (used with permission from publisher).

1.3 Modeling Patient Survival

With any prognostic or response assessment model, there needs to be a robust framework for estimating survival itself and correlation between variables (sometimes called covariates) and survival. A brief introduction to some relevant concepts in survival statistics follows. Ideally, patient time-to-event data is measured from a set landmark time (i.e., from diagnosis or treatment initiation) until the observed event of interest. This event may consist of disease progression, disease relapse, or ultimately death. The question of interest may simply be "what is is the median survival?" or may be more complicated, such as "is the median survival significantly different between males and females for patients with mesothelioma?" In either case, the question is fairly easy to answer if all events are known to occur.

However, the event of interest is not always observed, for instance when a patient is "lost" to follow-up. All that is known in these cases is that the event of interest is *at least* a certain value, but the true value is unknown. This general loss of information is called "censoring," and specifically "right censoring" when the event is only known to occur after a certain duration [61, 62]. It is possible for time-to-censoring to be dependent on time-to-event: healthy patients would usually live longer, giving them more opportunity to move away from the study center, leading to loss in follow-up. However, censoring is usually assumed to be independent of the event of interest. Additionally, censoring can occur because of administrative issues, for instance if patients are still alive at the time that final data analysis is performed.

Estimates of patient survival time distributions are commonly obtained using the Kaplan-Meier method of analysis [63]. This method uses conditional probabilities over time to produce the maximum likelihood estimate (MLE) of patient survival. The differences between Kaplan-Meier estimates of survival for unique groups can be tested using log-rank analysis, which orders the event times between two or more groups [64]. Typically, the null hypothesis for the log-rank test is that there is no difference in the distribution of event times between the groups, and this hypothesis is tested using the rank score of each group compared with the others [65]. The log-rank test is

optimally powered for the proportional hazards (PH) case, where the ratio of hazard rates between groups is constant over time (i.e., the hazard rates of distinct groups are proportional). The hazard rate $h(t)$ is interpreted as the instantaneous probability of the event to occur, given that it has not yet occurred, and is related to the survival function $S(t)$ via

$$S(t) = \exp\left[-\int_0^t h\left(t'\right)\,\mathrm{d}t'\right]. \tag{1.1}$$

The most common method used to model patient survival and assess the correlation between covariates and survival is Cox proportional hazards (PH) modeling [61,66]. The standard PH survival model assumes that the effect of any given covariate is fixed across time, and $h(t)$ takes the form

$$h\left(t \mid \vec{Z}\right) = h_0(t)\exp\left(\vec{\beta}\cdot\vec{Z}\right), \tag{1.2}$$

where $h_0(t)$ is an arbitrary "baseline" hazard function, $\vec{Z}$ is the vector of covariates, and $\vec{\beta}$ is the vector of regression coefficients (to be estimated). By estimating β and its confidence interval for each covariate individually (in a univariate model), we can assess whether any covariate is significantly associated with patient survival (i.e., β different from zero). With this standard approach, the covariate vector $\vec{Z}$ is assumed to be fixed across time and does not allow for covariates that change with time.

The incorporation of time-varying covariates into a survival model is not trivial, since the fitting form of the Cox PH model changes. The partial likelihood function (used to estimate the covariate effects in the same manner as the typical likelihood function) for the model becomes

$$L_P = \prod_{failure\,times\,j}\left(\frac{\exp\left(\vec{\beta}\cdot\vec{Z}_j\left(t_j\right)\right)}{\sum_{i\in R_j}\exp\left(\vec{\beta}\cdot\vec{Z}_i\left(t_j\right)\right)}\right), \tag{1.3}$$

where now the covariate vector $\vec{Z}(t)$ can change with time. The term R_j represents the patients still at risk at the j^{th} failure time (i.e., still alive and on-study), and in a typical Cox PH model

with fixed covariates, risk sets at later time points are subsets of earlier risk sets since patients can only leave the risk set, not re-enter. For patients who endure from one risk set to the next, their $\vec{\beta} \cdot \vec{Z}_i$ value does not change. In equation 1.3, though, there is additional computational complexity since the risk set R_j must be entirely recalculated at each unique failure time to properly account for covariates that change over time. While this extension of the PH model is not overly common, it is readily available in many statistical software packages [67]. There are multiple ways to report the findings of a survival model; the regression coefficients $\vec{\beta}$ and their confidence intervals can be tabulated, or the model can be depicted graphically using a nomogram. For examples, see [34, 38].

In order to assess the performance of a survival model, we need an appropriate metric. Receiver operating characteristic (ROC) analysis provides an interesting framework to assess survival curves. ROC analysis can be used when the data consist of an ordinal-scale predictor for a binomial outcome of interest and the true status of that outcome. Then, the sensitivity and specificity for predicting the outcome of interest can be calculated for a whole range of possible cut points in our predictor [68–70]. Plotting these sensitivity/specificity pairs on the typical ROC axes and calculating the area under the curve (AUC) gives one metric for how well the predictor performs.

In survival analysis models, at each time t, each patient will have a model-predicted probability of survival, and of course each patient will have survived to time t or will have experienced the event of interest by time t. Therefore, treating our group of model-predicted survival probabilities for our patients as our set of "predictors," we can come up with an ROC curve for our model at each time where a prediction is made by the model. For instance, at 6 months, the model will predict that some patients have a 90% chance of survival, some have a 70% chance of survival, etc., and by sweeping the decision threshold ("anyone with > 80% predicted survival rate will be called 'alive'...now which patients were actually alive at 6 months?"), we can construct the ROC curve. For a model using time-independent covariates, the predicted survival probability for any patient will change through time, but the ordering of the predicted survival probabilities among patients will remain constant due to the proportional hazards model. Combining equations 1.1 and

1.2 for a time-independent Cox PH model, it can be shown that

$$S(t) = S_0(t)^{\exp\left(\vec{\beta}\cdot\vec{Z}\right)}, \tag{1.4}$$

where $S_0(t)$ is calculated from equation 1.1 using $h_0(t)$ as described above. Therefore, predicted survival probability changes over time, but the monotonicity between predicted survival and a given patient's $\vec{\beta}\cdot\vec{Z}$ is preserved.

For these models with covariates fixed across time, the same ROC curve would be obtained at every time t, and Harrell's C statistic is equivalent to the non-parametric area under this single ROC curve [71]. However, if the covariates are allowed to vary with time, the rank ordering of predicted survival probabilities amongst patients can vary as well, and there will exist a unique ROC curve for model performance at each time. Each of these ROC curves will have a corresponding AUC, and plotting this AUC as a function of time gives us a general sense of our model performance. This is precisely the motivation for Heagerty's time-dependent AUC and C^τ [72], where we calculate our final performance metric as

$$C^\tau = \int_0^\tau \mathrm{AUC}(t)\cdot w^\tau(t)\,\mathrm{d}t, \tag{1.5}$$

where AUC(t) is the AUC as a function of time explained above, τ is the follow-up period of interest, and w^τ is a weight function. The problem has been considered by others, but Heagerty's metric is available in R [73, 74]. Harrell's C can also be used to assess the performance of a single factor in predicting patient survival outside of a Cox model.

1.4 Dynamic Imaging

While standard CT scans provide structural information based on differing tissue densities, the similarity between surrounding soft tissue density and MPM density on CT scans complicates the accurate identification of disease and evaluation of tumor growth. The use of intravenous

iodinated contrast media can lead to an increase in the differentiation between tissue types, but there is still considerable overlap between MPM and other thoracic tissues [24]. In a traditional contrast-enhanced scan, an amount of contrast appropriate for the patient (usually 90-120 mL of contrast with an iodine concentration of 350 mg/mL) is injected intravenously, and the CT scan is acquired after an amount of time that allows for distribution of the contrast media throughout the body (usually 30–60 seconds). Such a scan still results in a single timepoint image, or snapshot, but the resulting scan contains information related to both anatomy and vascularity.

Still more information related to vascularization and functional physiology can be discovered through the use of perfusion imaging, often called dynamic contrast-enhanced (DCE) imaging. The main difference with DCE imaging is that instead of acquiring a single timepoint scan after a set delay from initiation of injection, scans are repeatedly captured during (and potentially following) the intravenous injection of the contrast media. In this way, DCE imaging adds a temporal component to an otherwise structural scan. Because the change in HU value for a given voxel is directly proportional to the contrast of iodine in that voxel, blood flow to different tissues can be tracked over time from the temporal scan information. If a single timepoint contrast-enhanced scan gives information about relative vascularity between various soft tissues, a DCE scan gives information about how fast the blood is flowing to the tissues of interest, how much blood is flowing through those tissues, and how long the blood takes to travel through those tissues.

The application of a DCE-CT protocol will create data maps for hemodynamic properties that may demonstrate increased separation between MPM and surrounding soft tissues. The same hemodynamic maps of blood flow and vascularity have also been shown to serve as useful oncologic biomarkers [75]. In a study using DCE-MRI to image MPM patients, Giesel *et al.* [76] reported that a disease perfusion parameter (k_{ep}) was clearly associated with patient survival, and two other perfusion values, taken together, differentiate MPM tissue from surrounding soft tissues. However, the general application of DCE-MRI for MPM is limited by the presence of cardiac and respiratory motion. It is therefore reasonable to believe that DCE-CT will produce functional data

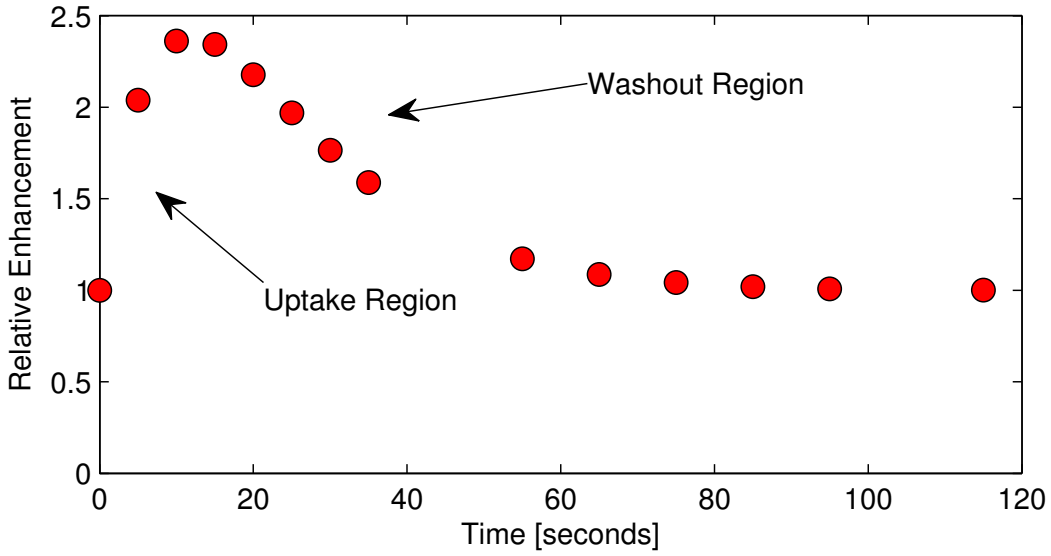


Figure 1.5: A hypothetical series of data from a DCE-CT scan. The individual uptake and washout characteristics of a single voxel can be observed from enhancement due to perfusion of iodinated contrast media to the tissues of interest.

maps that hold prognostic significance and help to differentiate MPM from thoracic soft tissue in a manner similar to DCE-MRI imaging. And while DCE-CT uses ionizing radiation to generate scan data, the scans do not suffer from the MRI artifacts mentioned previously, and the DCE-CT scan can be integrated nicely with the current CT scanning already being performed on MPM patients.

A DCE-CT scan begins when a bolus of iodinated contrast media is injected venously via a power injector. The contrast agent enters the circulatory system and perfuses to different tissue types according to their vascularity and blood flow, where the increased attenuation of iodine will present with increasing HU value [77]. The degree of tissue enhancement depends on the contrast media concentration and local hemodynamics reflecting tissue vascularity and physiology [78]. During the DCE-CT scan, a small section of relevant patient anatomy (limited axially by the extent of the collimated CT beam) is repeatedly imaged, and the uptake and eventual washout characteristics of the iodinated contrast can be quantified by tracking voxel-by-voxel HU values over time. A hypothetical series of data for one voxel can be seen in Figure 1.5.

One uptake curve is obtained for every voxel in the three-dimensional image, and various

oncologic biomarkers can be derived from these curves. The Philips scanner used in this research will determine: perfusion, tissue peak enhancement, time to peak enhancement, and mean transit time [77, 79]. Perfusion measures the flow rate through the vasculature of the structure and is an oncologic marker for tumor vascularity and grade. Perfusion (along with the other parameters) is calculated according to the "slope method" of deconvolution analysis, and is calculated as the ratio of the maximum slope in the tissue uptake curve to the peak arterial enhancement, or

$$\text{Perfusion} = \frac{\max\left(\frac{\mathrm{d}HU(t)_{tissue}}{\mathrm{d}t}\right)}{\max\left(HU(t)_{arterial} - HU(0)_{arterial}\right)}. \tag{1.6}$$

Peak enhancement is the maximum increase in tissue density over baseline, or

$$\text{Peak Enhancement} = \max\left(HU(t)_{tissue} - HU(0)_{tissue}\right), \tag{1.7}$$

and is an oncologic marker of tissue blood volume. The marker commonly called "blood volume" is simply a normalized version of peak enhancement and is calculated by the slope method as

$$\text{Blood Volume} = \frac{\max\left(HU(t)_{tissue} - HU(0)_{tissue}\right)}{\max\left(HU(t)_{arterial} - HU(0)_{arterial}\right)}. \tag{1.8}$$

The time to peak enhancement is a measure of the time for contrast to flow from a major arterial vessel to the tissue, or

$$\text{Time to Peak} = \arg\max\left(HU(t)_{tissue}\right) - t_{\text{arterial arrival}}, \tag{1.9}$$

and is an oncologic marker of pressure (where $t_{\text{arterial arrival}}$ is the time of initial contrast enhancement in the arterial input). Mean transit time is the time taken for contrast to travel from artery to vein, calculated as the full-width at half-maximum (FWHM) of the tissue uptake curve over

baseline, or

$$\text{Mean Transit Time} = \text{FWHM}\left(HU(t)_{tissue} - HU(0)_{tissue}\right), \tag{1.10}$$

and is an oncologic marker of perfusion pressure.

To our knowledge, DCE-CT has been used for imaging of MPM in only a single study conducted by Meijerink *et al.* [80]. This study investigated the use of DCE-CT to study the hemodynamic response of tumors of the thorax and abdomen. The patient cohort consisted of 16 patients, including only two patients with confirmed MPM. DCE-CT data demonstrated a substantial decrease in tumor perfusion values for both MPM patients after treatment with anti-angiogenic VEGF inhibitors. This supported the assertion that DCE-CT is able to measure hemodynamic function in MPM patients.

In the current climate of caution surrounding all things radiological, the radiation dose from DCE-CT has come under scrutiny [81, 82]. Even before recent accidents were made public, researchers questioned the necessity of using DCE-CT when DCE-MRI and PET imaging were already available [23]. However, comparing the effective whole-body doses of FDG-PET imaging and DCE-CT imaging of the lung, a typical FDG-PET body scan can contribute on average 5–10 times (and up to 20 times) the effective dose of a DCE-CT scan [83, 84]. Therefore, DCE-CT may be favorable to FDG-PET imaging for the limited task of assessing local tumor physiology if DCE-CT can reveal functional tumor characteristics that are similar in their prognostic significance to FDG-PET measurements. FDG-PET imaging will always maintain certain advantages over DCE imaging, however, since an FDG-PET scan yields data for the *entire* tumor extent as well as distant data from nodal and metastatic involvement.

1.5 Outline

The following dissertation will describe a series of projects with the overall goal of redefining radiologic response for patients with malignant pleural mesothelioma. Response assessment, a key

component of clinical trials for MPM, is currently performed using linear measurements of disease burden. Many phase II clinical trials, where radiologic response assessment plays a crucial role, show promising results, only to disappoint during the phase III trial (not only for MPM) [40, 41]. Therefore, it is reasonable to question the validity of the current response metric widely used in phase II trials, namely the RECIST system.

As a coherent progression, this work will extend the definition of radiologic response for MPM patients from one to three dimensions, both in the context of a discrete classification system and as a continuous prognostic model. In Chapter 2, I will optimize the discrete response classification criteria for patients with MPM using the current linear (1D) measurement technique. Image-based response assessment is often used as a surrogate for patient benefit in clinical trials, and therefore the most useful (and relevant) response classification criteria would be those developed to maximize the surrogacy of the assessment metric for meaningful outcomes such as survival. Such classification criteria would be better suited to the specific morphology of mesothelioma and would provide more realistic markers of treatment efficacy, all while preserving the simplicity of the current linear measurement protocol. I hypothesize that without changing the measurement technique, a new set of classification criteria can be obtained to more appropriately cluster patients into response categories than the current -30%/+20% standard RECIST criteria.

In Chapter 3, I will investigate the potential use of manual area (2D) contours as a means of response assessment in MPM patients. Area contours may serve as a useful "middle ground" for response assessment in MPM. The area contouring process may be better suited to capture morphological tumor changes in MPM, and by summing the area measurements from a limited number of axial sections, observers are spared the time and effort required to contour complete volumes of disease. However, the additional "degrees of freedom" inherent in the area contouring process may lead to high levels of inter-observer variability in area measurements. I hypothesize that while perhaps able to better capture changes in MPM burden, the inter-observer variability between manual area contours will exceed any reasonable response criteria and is therefore too

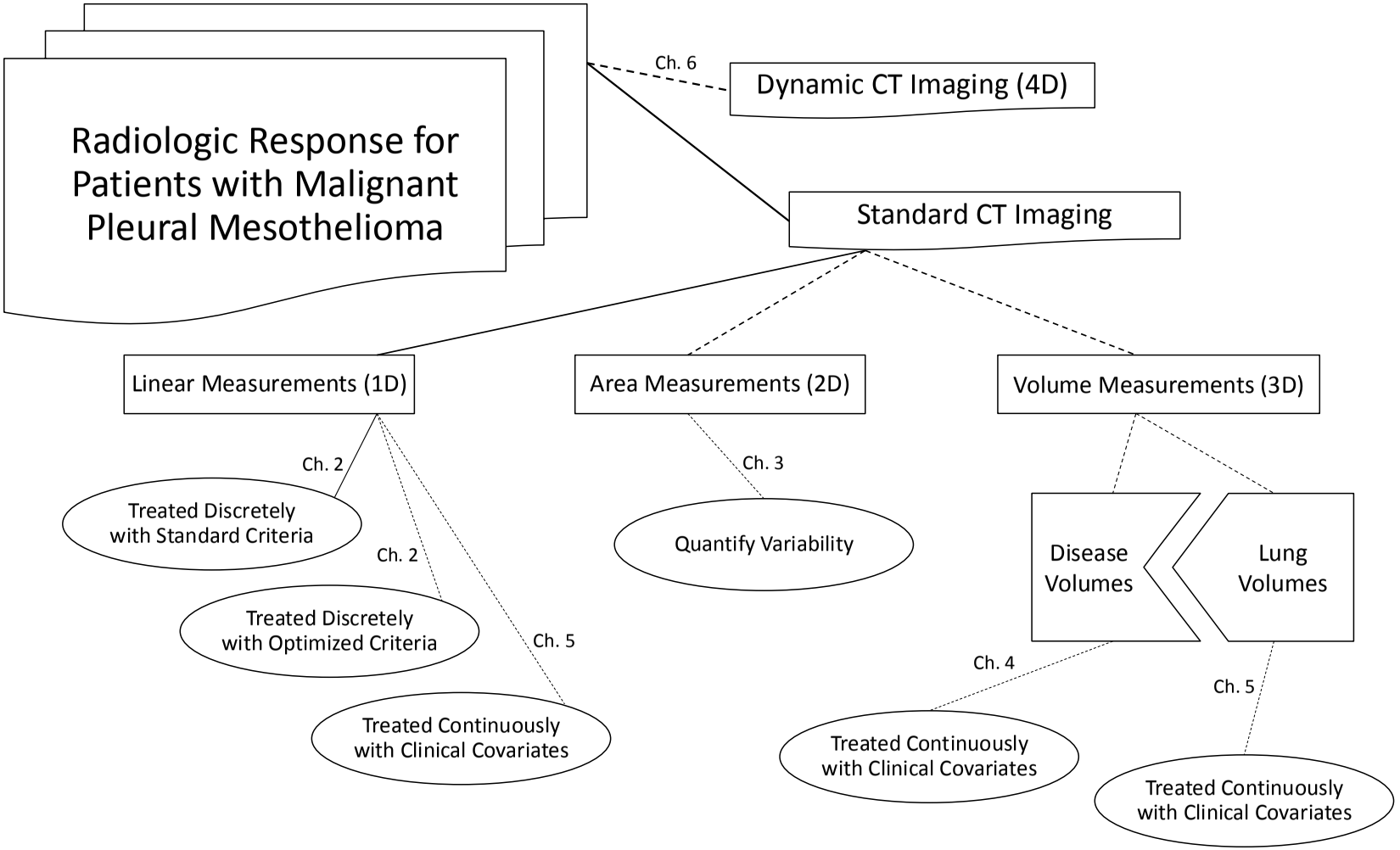


Figure 1.6: Graphical depiction of the scope of work in this dissertation.

large to allow for robust response assessment.

In Chapter 4, I will incorporate time-changing radiologic measurements of disease volume, as well as other clinical covariates, into a survival model for patients with MPM. Disease volumes are the most direct and meaningful measurements of tumor burden. This study will serve as the most "complete" survival model for MPM in this dissertation. I hypothesize that the addition of time-changing radiologic measurements to survival models with other clinical covariates will prove significant for the prediction of patient survival.

In Chapter 5, I will assess the prognostic significance of changes in automatically segmented lung volume for use as a novel response assessment metric. Lung volume serves as a physiological correlate of disease volume for pleural diseases such as MPM. Because the thoracic space is approximately bounded by the ribcage, it is reasonable to expect changes in lung volume to be complementary to changes in disease volume, and therefore that changes in lung volume may be used as a surrogate for changes in disease volume. Survival models will be constructed in the same manner as in Chapter 4, and models using disease volumes will be compared to models using either lung volumes or linear measurements. Here, I hypothesize that changes in lung volumes will correlate with changes in disease volume, and that changes in lung volume will be significantly associated with patient survival. The complete survival models built using the three measurement techniques will be compared and contrasted.

In Chapter 6, I will investigate DCE-CT as a functional imaging tool for MPM. The DCE-CT imaging protocol is an experimental method allowing for the integration of DCE imaging with clinically standard contrast-enhanced imaging of the full thorax. In this pilot study, changes in DCE-CT metrics will be correlated with changes in disease burden (both linear and volumetric measurements) for patients undergoing treatment as well as for patients on observation. Patterns of interest will be reported for potential investigation in future response assessment trials with DCE-CT in MPM.

Finally, in Chapter 7 I will provide a summary of the relevant conclusions from the separate in-

vestigations, as well as make recommendations regarding future directions of study. The studies in this dissertation form part of the answer to the question "Can we successfully and effectively redefine radiologic response for patients with malignant pleural mesothelioma?" Hardly any question can ever be considered truly answered in the sciences, and this dissertation will hopefully serve its role as a small impetus of change in the small world of response assessment for malignant pleural mesothelioma.

CHAPTER 2

OPTIMIZATION OF RESPONSE CLASSIFICATION CRITERIA

"We cannot solve our problems with the same thinking we used when we created them." – Albert Einstein

2.1 Introduction

As explained in section 1.2 above, the history of the Response Evaluation Criteria in Solid Tumors (RECIST) classification criteria casts doubt on the applicability of such criteria for classification of response in a disease so typically aspherical as mesothelioma. Indeed, alternate theoretical response classification criteria for mesothelioma have been derived from geometric models other than spheres, such as the lens, crescent, or annulus [51]. On the whole, these geometric models indicate that uni-dimensional measurements acquired according to the modified RECIST guidelines would be better classified with criteria where the definition of stable disease was more broad. Typically, if the uni-dimensional measurement increased by more than 20% from the minimum of previous measurements (nadir), the patient was classified as having progressive disease (PD), and if the measurement decreased by 30% or more from the baseline measurement, the patient was classified as partially responsive (PR). Stable disease (SD) was defined for measurement changes lying between the previously mentioned classification groups. For the alternate geometry models, the current short-axis measurement criteria of -30%/+20% would be replaced by -66%/+74% for the crescent geometry, and for the lens and annulus geometries, the criteria would be -52%/+45% and -68%/+100%, respectively.

While these models raise important issues, they are still theoretical derivations. Image-based response assessment is often used as a surrogate for patient benefit in clinical trials, and therefore the most useful (and relevant) response classification criteria would be those developed to maximize the surrogacy of the assessment metric for meaningful outcomes such as survival. Such

classification criteria would be better suited to the specific morphology of mesothelioma and would provide more realistic markers of treatment efficacy.

The purpose of this specific study is two-fold: first, to assess the validity of existing response classification criteria in malignant pleural mesothelioma (MPM), and second, to optimize disease-specific response classification criteria by maximizing the correlation between response assessment and overall survival in patients with MPM treated with chemotherapy. The main hypothesis for this study is that without changing the measurement technique, a new set of classification criteria can be obtained to more appropriately cluster patients into response categories.

2.2 Patients and Methods

2.2.1 Patient Cohort

The patient cohort for this study was a subset of a database collected retrospectively from Sir Charles Gairdner Hospital in Perth, Western Australia, Australia, in collaboration with Anna Nowak, M.D., Ph.D. of Sir Charles Gairdner Hospital and the University of Western Australia. The patients were part of a prospective study involving fluorodeoxyglucose positron emission tomography (FDG-PET) and CT imaging of MPM [38]. All patients were over 18 years old with histologically or cytologically confirmed MPM and had not received prior chemotherapy or radiotherapy. Original patient accrual occurred from late 2003 to 2010, and the original study was approved by the local institutional Human Research Ethics Committee at Sir Charles Gairdner Hospital, with patients providing written informed consent. The retrospective analysis of the HIPAA-compliant data was approved by both the originating institution's Human Research Ethics Committee and the Institutional Review Board at The University of Chicago, where the analysis was performed. Because the original study was not a research treatment study, patients were treated as clinically indicated. Initially, combination chemotherapy consisted of cisplatin and gemcitabine, and later, when it became available at the original study institution, cisplatin and pemetrexed. Palliative

radiotherapy was used when indicated.

The original full database consists of 129 patients. However, for the studies in this dissertation, patients were only included where imaging and clinical data existed for first-line chemotherapy (i.e., not only during secondary salvage therapy). Furthermore, patients were only investigated when serial imaging studies existed in a digital format. These restrictions reduced the number of eligible patients to 97. Summary descriptive information for the 97 patients eligible for consideration in this dissertation is given in Table 2.1.

Table 2.1: Description of the full eligible patient cohort, consisting of 97 patients. The specific patient cohorts used in the studies of this dissertation are subsets of this cohort.

Characteristic	Summary
Sex:	
Male	n = 84
Female	n = 13
Age at Diagnosis:	
Median	66 years
Range	41–80 years
Chemotherapy:	
Carboplatin/Pemetrexed	n = 7
Cisplatin/Pemetrexed	n = 46
Cisplatin/Gemcitabine	n = 44
Histology:	
Epithelioid	n = 72
Sarcomatoid	n = 8
Biphasic	n = 17

(continued on next page)

(Table 2.1, continued from previous page)

T Stage:		
	T1	n = 21
	T2	n = 26
	T3	n = 32
	T4	n = 18
N Stage:		
	N0	n = 29
	N1	n = 3
	N2	n = 49
	N3	n = 16
M Stage:		
	M0	n = 83
	M1	n = 14
IMIG Stage:		
	I	n = 14
	II	n = 6
	III	n = 45
	IV	n = 32
Known Asbestos Exposure:		
	Yes	n = 91
	No	n = 6
Chest Pain:		
	Yes	n = 61
	No	n = 36

(continued on next page)

(Table 2.1, continued from previous page)

Shortness of Breath:		
	Yes	n = 79
	No	n = 18
ECOG Performance Status:		
	0	n = 44
	1	n = 48
	2	n = 5
Talc Pleurodesis:		
	Yes	n = 36
	No	n = 61
Weight:		
	Median	76 kg
	Range	52–121 kg
Height:		
	Median	173 cm
	Range	155–190 cm
Smoking Status:		
	Never	n = 43
	Past	n = 47
	Present	n = 7
Pleurectomy/Decortication:		
	Yes	n = 4
	No	n = 93

For inclusion in this specific study, patients were required to have available modified RECIST

tumor thickness measurements at baseline (prior to beginning chemotherapy) and for one or more follow-up scans during chemotherapy. This requirement did not always imply the availability of the *complete* digital CT scans for the dates of measurement, since original baseline measurements were occasionally made on scans from an outside institution, where scans were transmitted as digitized films. Additionally, patients were required to have measurable disease at therapy baseline, which excluded some patients whose disease had been surgically removed prior to treatment with chemotherapy. These new constraints reduced the eligible patient cohort to 78 patients, and the summary description of these patients is given in Table 2.2. The remaining 19 patients were eligible for inclusion in the other studies in this dissertation.

Table 2.2: Description of the patient cohort used in this specific study, consisting of 78 of the original 97 patients. This specific patient cohort is a subset of the patients summarized in Table 2.1.

Characteristic	Summary
Sex:	
Male	n = 66
Female	n = 12
Age at Diagnosis:	
Median	66 years
Range	41–80 years
Chemotherapy:	
Carboplatin/Pemetrexed	n = 7
Cisplatin/Pemetrexed	n = 41
Cisplatin/Gemcitabine	n = 30

(continued on next page)

(Table 2.2, continued from previous page)

Histology:		
	Epithelioid	n = 56
	Sarcomatoid	n = 7
	Biphasic	n = 15
T Stage:		
	T1	n = 16
	T2	n = 18
	T3	n = 26
	T4	n = 18
N Stage:		
	N0	n = 22
	N1	n = 3
	N2	n = 38
	N3	n = 15
M Stage:		
	M0	n = 65
	M1	n = 13
IMIG Stage:		
	I	n = 11
	II	n = 3
	III	n = 34
	IV	n = 30

(continued on next page)

(Table 2.2, continued from previous page)

Known Asbestos Exposure:		
	Yes	n = 72
	No	n = 6
Chest Pain:		
	Yes	n = 49
	No	n = 29
Shortness of Breath:		
	Yes	n = 66
	No	n = 12
ECOG Performance Status:		
	0	n = 35
	1	n = 38
	2	n = 5
Talc Pleurodesis:		
	Yes	n = 32
	No	n = 46
Weight:		
	Median	75 kg
	Range	52–121 kg
Height:		
	Median	172 cm
	Range	155–190 cm

(continued on next page)

(Table 2.2, continued from previous page)

Smoking Status:	
Never	n = 36
Past	n = 37
Present	n = 5
Pleurectomy/Decortication:	
Yes	n = 1
No	n = 77

2.2.2 Imaging

Patients were imaged using helical CT up to one month prior to the first cycle of chemotherapy and throughout their treatment regimen (typically after the first cycle, then every two cycles thereafter). Images were reconstructed axially with 5-mm slices. CT staging was performed according to the Union for International Cancer Control (UICC) TNM staging system (2002). CT scans were staged by a thoracic radiologist or medical oncologist experienced in mesothelioma imaging, and tumor measurements were made according to the modified RECIST protocol on baseline and all follow-up scans [47]. Initially, radiologic response was classified according to the standard RECIST criteria, where partial response (PR) is a 30% reduction in tumor thickness over baseline, progressive disease (PD) is a tumor thickness increase of 20% over the nadir measurement, and stable disease (SD) is attributed to patients who failed to meet the criteria for either of the other categories.

There were a total of 275 CT scans in this study, with a median of four scans per patient (including baseline scans). Eleven patients had only a baseline scan with one follow-up scan, while 25 patients had three scans total, 32 patients had four scans total, and 10 patients had five scans total. The median duration between scans was 45 days.

2.2.3 Correlating Response with Survival

To measure the association between patient response classification and survival, a single response category must be assigned for each patient, since a response classification is actually assigned at each follow-up scan during the patient's treatment. This single response can be assigned in multiple ways; for instance, the "best response" for the patient achieved during some time interval (usually over the active treatment period) can be used, where PR is "better" than SD, and SD is better than PD. Alternatively, response at a predetermined follow-up time could be equally important. In this study, both the best response and response at the first follow-up scan were investigated.

If the response assessment system were a perfect correlate of survival (measured from diagnosis), patients labeled as PR would survive longer than those patients labeled as SD, and both groups would survive longer than patients labeled as PD. It is important to note that this is different from the groups having different and proportional hazard rates, which is the metric of interest in Cox proportional hazards modeling and log-rank testing [61, 65, 66]. When groups have different hazard rates, and the hazard ratios between groups are assumed to be independent of time (as they are in traditional Cox PH modeling), equation 1.4 reveals that the hazard ratio simply has an exponential effect on estimated survival. That is, unique and proportional hazard rates between groups *do not* imply that survival times of the different groups will be separated, but only that individuals in the different groups will experience the event of interest at different rates. This important distinction motivates a correlative metric choice other than the log-rank metric in this study, since surrogacy between response assessment and patient survival is most relevant when survival times are *separated* between groups.

The extent to which the desired trend holds true (PR survives longer than SD, SD survives longer than PD) can be measured quantitatively using Harrell's C statistic, an extension of Smith's P_k measure of rank correlation [71, 85]. C is scaled similarly to the standard area under the receiver operating characteristic (ROC) curve (AUC); a value of $C = 0.5$ is equivalent to classification by chance alone, whereas $C = 1.0$ would indicate perfect separation of response groups with respect to

subsequent survival times. According to Harrell, $C > 0.65$ indicates clinical utility, while $C > 0.80$ indicates high predictive accuracy [71]. The numerical value of C can be interpreted as the fraction of patient comparisons that would be "concordant." A concordant pair of data points exists when the patient with the higher response rating is known to survive at least as long as the patient with the lower response rating, which allows for the appropriate consideration of censoring.

For example, when $C = 0.70$, in 70% of the valid patient comparisons where patient B is classified into a "better" response category than patient A, patient B will be known to live at least as long as patient A. Graphical examples are shown in Figures 2.1 and 2.2. In Figure 2.1, the response category is actually anti-correlated with survival, which would be opposite the expected trend, and therefore we would expect C to be less than 0.5. In Figure 2.2, however, response categorization is nearly perfectly associated with survival, and therefore we expect C to be very nearly unity. All analyses were performed using the academic edition of Revolution R Enterprise (version 4.3, based on R version 2.12), and C was implemented in the R package "Hmisc" [86,87].

2.2.4 Optimization and Cross-Validation

To determine the optimal set of response classification criteria, the PR and PD cutoffs were varied in 1% increments (i.e., the PR cutoff was swept from -100% to 0% in 1% increments, and the PD cutoff was swept from 0% to 100% in 1% increments). For each possible pair of cutoff criteria, the correlation between response and survival was assessed to yield one value of C. By tabulating the values of C across all possible response criteria, the optimal pair of classification criteria was determined. These optimal points represent the classification criteria for which the correlation between response and patient survival is greatest. The criteria derived in this way, using all 78 patients, will be called the "full cohort criteria."

The optimization process requires validation, since the most optimal model from the full patient cohort has a strong tendency to yield an overly optimistic prediction rule with respect to predictions on *de novo* observations not involved in model building. A leave-one-out cross-validation

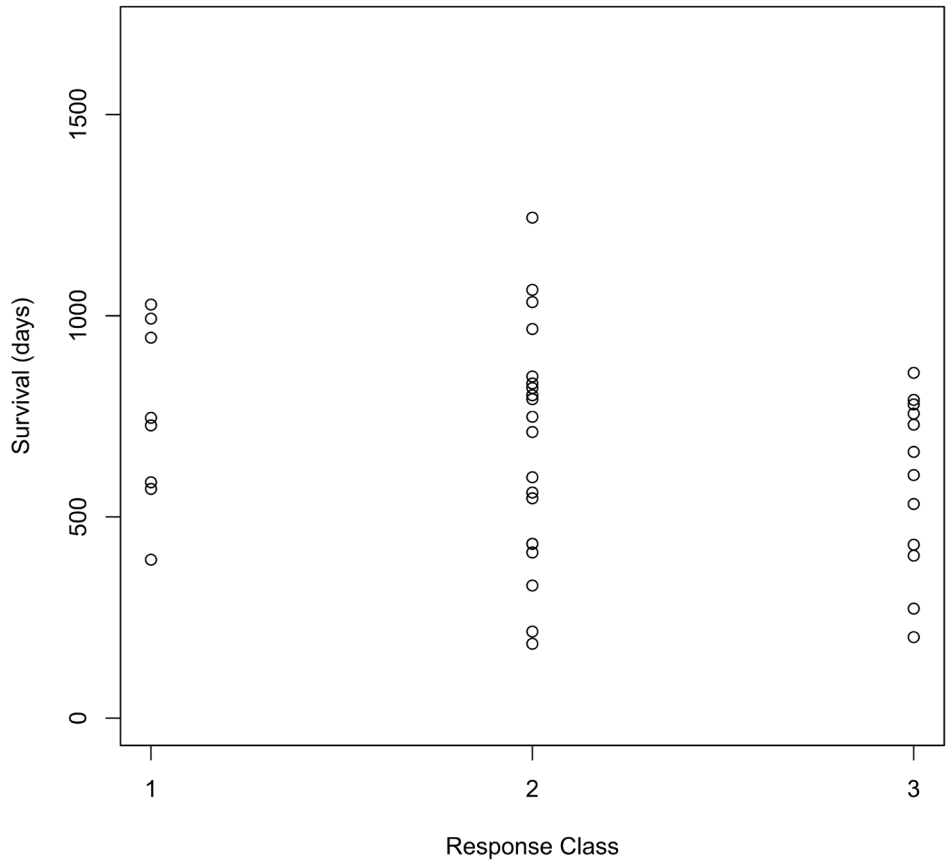


Figure 2.1: Hypothetical distribution of survival with response classification. For these simulated data, response groups 1, 2, and 3 would correspond to progressive disease, stable disease, and partial response, respectively. Using Harrell's C, the performance of these simulated data is $C = 0.40$.

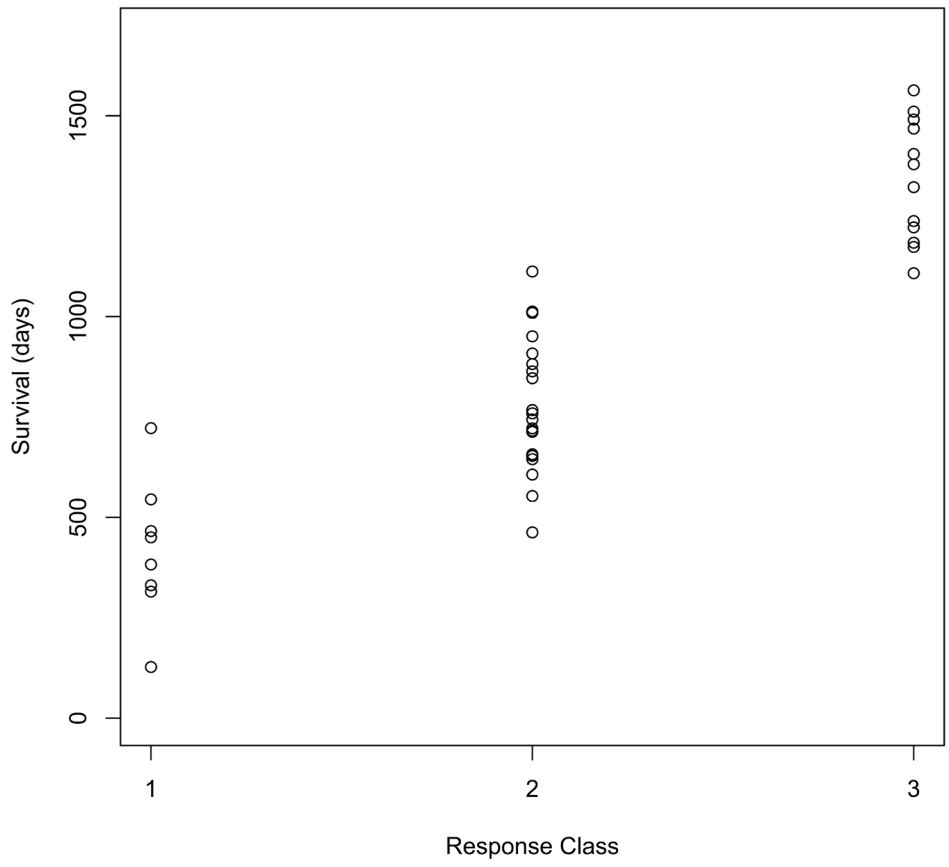


Figure 2.2: Hypothetical distribution of survival with response classification. For these simulated data, response groups 1, 2, and 3 would correspond to progressive disease, stable disease, and partial response, respectively. Using Harrell's C, the performance of these simulated data is $C = 0.98$.

(LOOCV) process may thus generate a more realistic value of C [71, 88]. Using LOOCV, each patient was excluded, one at a time, and the classification criteria were optimized using the other 77 patients. The optimized criteria from these 77 "training" patients were used to test the model on the 78th patient (who had been excluded from model optimization). This process continued until all 78 patients individually had been excluded from the optimization cohort. In this manner, all 78 patients were assigned a response category (for each of best response and first follow-up response), and a value of C correlating these LOOCV-based response categories to patient survival was calculated. LOOCV allows each patient to be assigned to a response category using criteria that were derived without knowledge of that particular patient's tumor measurement trajectory, and the LOOCV-based value of C is a better indicator of how well the reported classification criteria will perform for a new, previously unknown patient.

Two additional internal validation checks were also performed. First, to evaluate possible dependence of the classification rule and C derived to specific cases or subsets of cases, bootstrap samples from the entire cohort were generated and the optimization procedure repeated, followed by a descriptive summary of the cut-points and C values. Secondly, to evaluate the performance of the derived rule from the full cohort in hypothetical independent patient cohorts, the rule was applied to random bootstrap sub-cohorts and the performance summarized.

When point estimates, or individual values, of C are calculated, standard errors (SE) are also calculated for the metric. However, point estimates of the performance of the standard RECIST classification criteria, C_{std}, and the optimized classification criteria performance, C_{opt}, will necessarily be correlated for a given patient sample because of the inevitable overlap in response classification groups. Therefore, to compare differences in point estimates of C, one must account for this correlation, since an assumption of independence would result in an overly conservative p-value for the difference between C_{std} and C_{opt}. There are multiple methods one could use to account for this correlation. One approach would be the simple calculation of the ordinal correlation between the standard RECIST response category for each patient and the new response category

from the optimized classification criteria. Another approach would use the pseudo-values from a jackknife analysis to implicitly discover the correlation between performance metrics [89, 90]. In general, the variance of a difference between two statistical quantities **X** and **Y** is given by

$$\mathrm{Var}[\mathbf{X}-\mathbf{Y}] = \mathrm{Var}[\mathbf{X}] + \mathrm{Var}[\mathbf{Y}] - 2\,\mathrm{Corr}[\mathbf{X},\mathbf{Y}]\,\sqrt{\mathrm{Var}[\mathbf{X}]\,\mathrm{Var}[\mathbf{Y}]}, \tag{2.1}$$

where Corr denotes the correlation between the two variables. Since correlation is zero for independent variables, assuming that C_{std} is independent from C_{opt} would artificially inflate the variance of their difference. Using the jackknife method, we compute the pseudo-values of C_{std} and C_{opt} for each patient in the full patient cohort [88]. These pseudo-values behave as if they are independent, and we fit a mixed effects linear model to the data, similar to a method previously described for ROC curves [89, 90]. Using the estimated variance components from the linear model, we can derive the correlation between C_{std} and C_{opt}. Finally, after the correlation between point estimates of C was factored into the difference, C_{opt} was compared with C_{std} using a one-sided Z-test with $p < 0.05$ as the standard for statistical significance.

2.3 Results

2.3.1 Patients and Overall Survival

Median overall survival from diagnosis was 14.9 months. Of the 78 patients, there were 75 observed deaths, while three patients were lost to follow-up after a median follow-up of 35 months. The overall survival curve is shown in Figure 2.3. Using the standard RECIST classification criteria, the median survival for best response PD, SD, and PR was 11.5 months, 11.6 months, and 23.0 months, respectively. Figure 2.4 shows group survival curves for best response using the standard RECIST classification criteria.

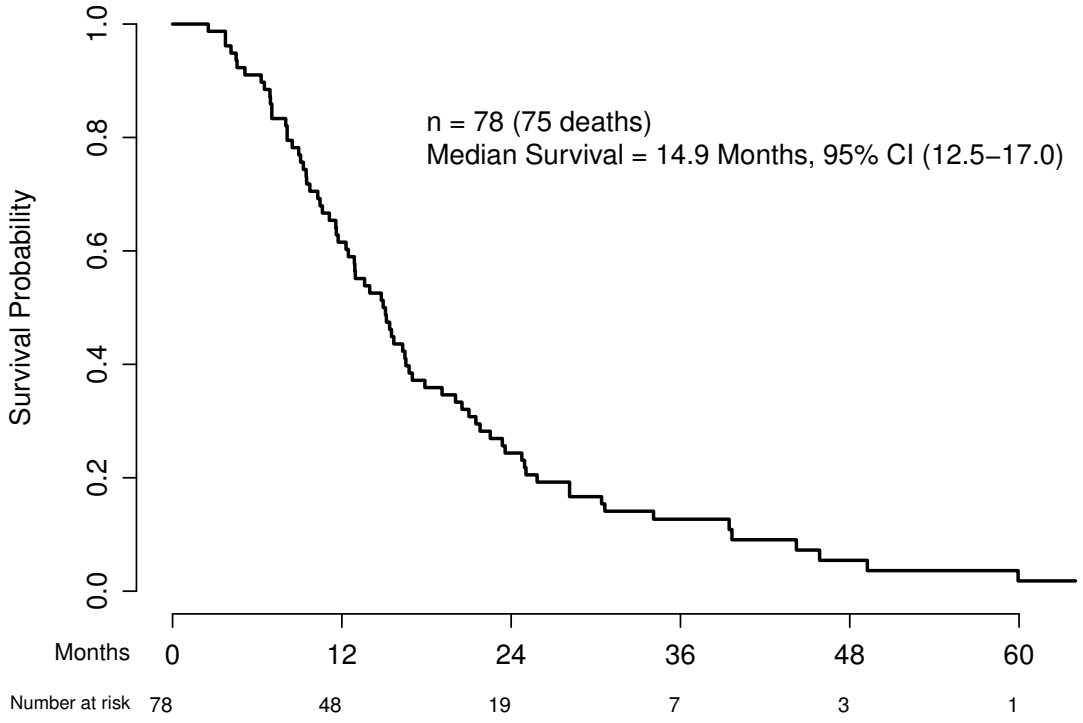


Figure 2.3: Overall survival curve for patient cohort.

2.3.2 Optimization of Classification Criteria

Using the standard RECIST classification criteria of -30% for PR and +20% for PD, the correlation between best response and overall survival was $C_{std}^{best} = 0.778$ with an SE of 0.048. The correlation between first follow-up response for each patient and overall survival was $C_{std}^{first} = 0.655$ with an SE of 0.054. After optimization, the new classification criteria derived from the full cohort were -64% for PR and +50% for PD. Optimizing the correlation between response classification and survival resulted in identical criteria using both the best response and first follow-up response per patient. The performance of these full cohort criteria using the best response per patient was $C_{opt}^{best} = 0.855$ with an SE of 0.045, and using the first follow-up response per patient, the performance was $C_{opt}^{first} = 0.932$ with an SE of 0.029. These values are summarized, along with their

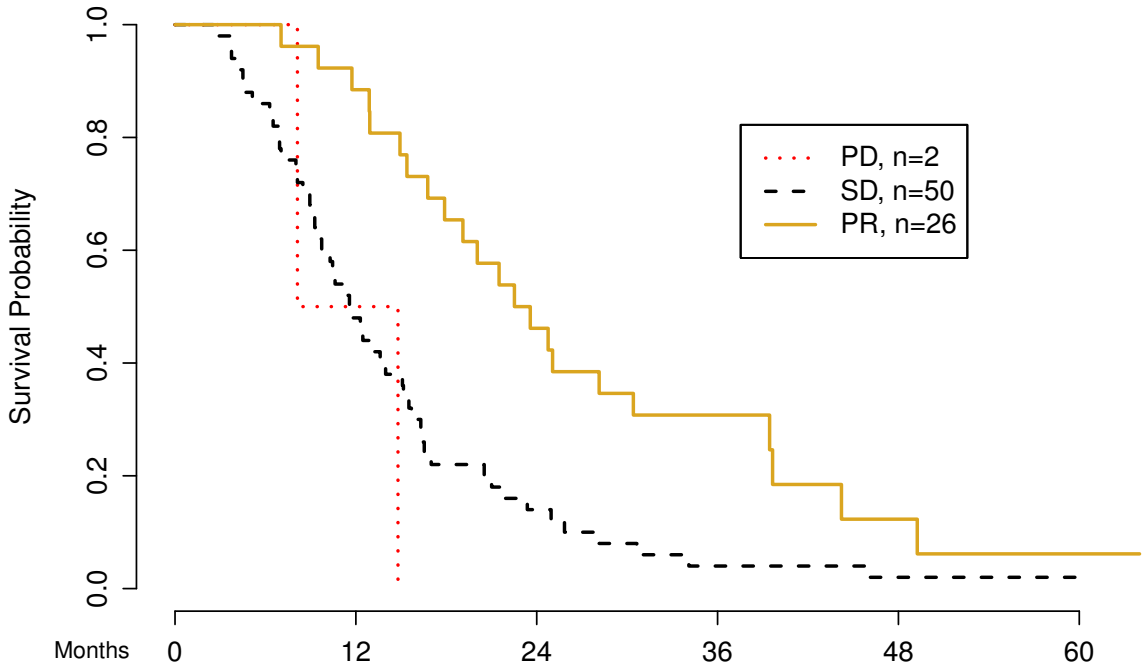


Figure 2.4: Overall survival from diagnosis by response category, using the standard RECIST -30%/+20% classification criteria and each patient's best response (progressive disease, PD; stable disease, SD; partial response, PR).

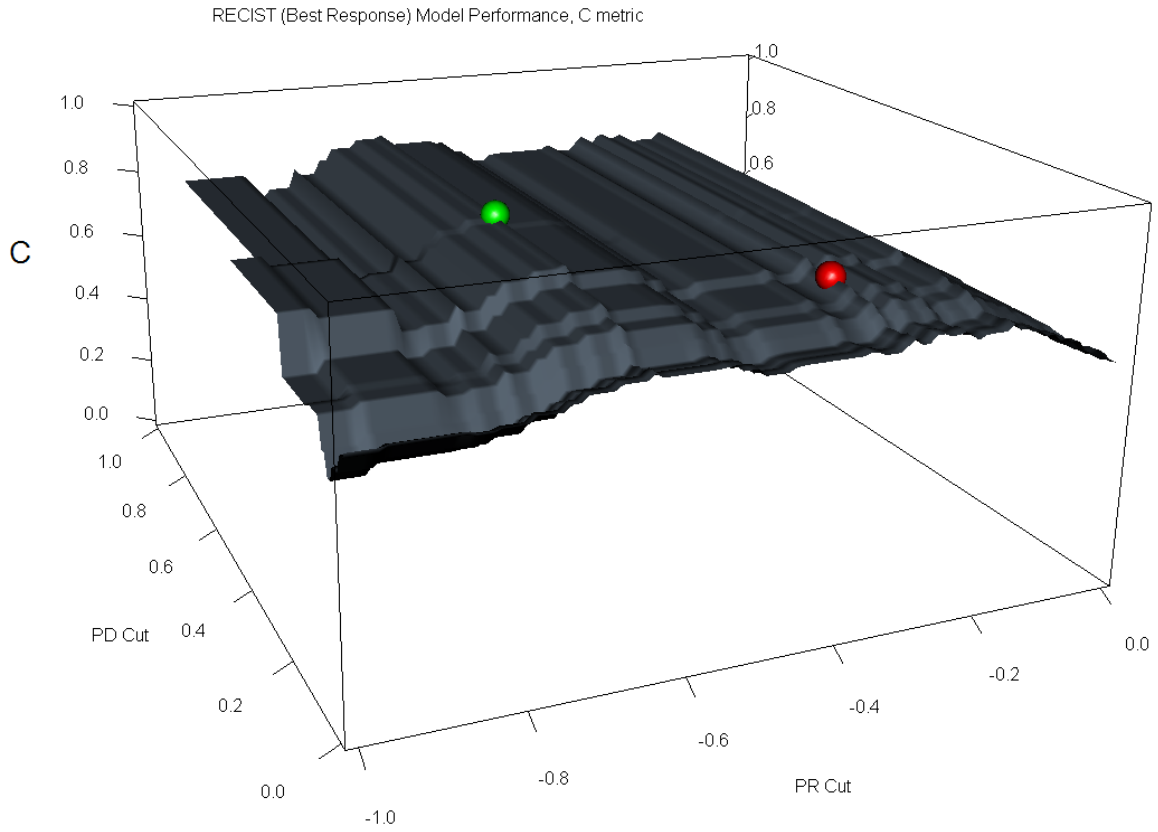


Figure 2.5: Surface plot showing the performance of the response classification criteria for various values of the PR and PD cut-points. Harrell's *C* is calculated at each set of possible classification criteria (PR from -100% to 0%, PD from +0% to +100%). The red dot indicates the standard RECIST criteria, and the green dot indicates the optimized -64%/+50% criteria.

p-values comparing optimized performance with the standard RECIST classification criteria performance, in Table 2.3. A surface plot is shown in Figure 2.5 to visualize the response model performance for various values of the classification criteria. From the surface plot, it can be seen that the optimized model performance exists on a fairly broad region and not on a singular prominent maximum.

Figure 2.6 plots the best response classification for each patient against overall survival from diagnosis using both the standard RECIST classification criteria and the optimized criteria. It can be seen that after optimization, the classification criteria group patients into only two response categories; the two patients originally classified as PD are now included in the SD category. Fur-

	Classification Criteria	C	Standard Error	p-value
Best response	Standard RECIST (-30%/+20%)	0.778	0.048	–
	Optimized (-64%/+50%)	0.855	0.045	0.039
	Cross Validation	0.829	0.043	0.121
First follow-up response	Standard RECIST (-30%/+20%)	0.655	0.054	–
	Optimized (-64%/+50%)	0.932	0.029	<0.001
	Cross Validation	0.872	0.049	<0.001

Table 2.3: Correlation scores between patient response and overall survival from diagnosis. All p-values are calculated with reference to the appropriate standard RECIST classification criteria performance (either best response or first follow-up response per patient) and properly account for correlation between values of C.

thermore, many of the patients originally classified as PR but having short survival durations are now included in the SD category. Figure 2.7 plots survival curves for the best response categories using the optimized criteria, where the median survival for best response SD and PR was 12.9 months and 24.8 months, respectively.

Table 2.4 shows a cross-tabulation of how patients are categorized using the standard RECIST and optimized classification criteria for both the best response and first follow-up response. For best response, 17 patients (22%) changed classification categories between the standard RECIST criteria and optimized criteria, and for first follow-up response, 10 patients (13%) changed classification categories.

		Optimized Classification Criteria		
		PR	SD	PD
Standard Classification Criteria	PR	11	15	0
	SD	0	50	0
	PD	0	2	0

(a)

		Optimized Classification Criteria		
		PR	SD	PD
Standard Classification Criteria	PR	1	8	0
	SD	0	67	0
	PD	0	2	0

(b)

Table 2.4: Number of patients in the different response categories using the standard RECIST classification criteria and the optimized -64%/+50% classification criteria. Response classified according to best response is shown in Table 2.4a, while response classified according to first follow-up response is shown in Table 2.4b.

2.3.3 Cross-Validation of Classification Criteria

As indicated in section 2.2.4, cross-validation of the optimized classification criteria leads to a more realistic value of model performance, C_{cv}. Correlating each patient's cross-validated best response with overall survival, a performance of $C_{cv}^{best} = 0.829$ with an SE of 0.043 was achieved. When the cross-validated first follow-up response was correlated with overall survival, the model performance was $C_{cv}^{first} = 0.872$ with an SE of 0.049. The LOOCV scheme is more a validation of the optimization process than any one set of optimized criteria, and therefore these C metrics are more realistic estimates of performance without the bias of training and testing a model on the same patient cohort. These C values, along with p-values comparing cross-validated performance with the standard RECIST performance, are summarized in Table 2.3.

From the first bootstrap internal validation (where the classification criteria were allowed to vary with each independent bootstrap patient sample), the criteria selected for each independent bootstrap sample are summarized as follows. The PR cut-point had a median value of -64%, with a mode of -64% and a mean of -67%, and the PD cut-point had a median value of +50%, with a mode

of +50% and a mean of +36%. In the second bootstrap internal validation, where the classification criteria were fixed at -64%/+50%, the mean performance across independent bootstrap samples was $C^{best}_{boot,opt} = 0.852$ with an SE of 0.047. For the same independent bootstrap samples, the mean performance of the standard RECIST criteria was $C^{best}_{boot,std} = 0.778$ with an SE of 0.050. A comparison of these bootstrap performance values and their respective standard errors to the values in Table 2.3 reveals them to be quite similar.

2.4 Discussion

In order to assess patient response to therapy, clinicians have come to rely on image-based measures of tumor burden as a surrogate for "true" patient benefit (i.e., reduced symptom burden or time until a defined event such as death). One common method for image-based assessment is the RECIST paradigm of linear measurements and response classification criteria. While the specific technique used to acquire tumor measurements has been defined in a specific sense for patients with MPM (modified RECIST), the response classification criteria for MPM patients are the same cut-points used for all tumors based on standard RECIST, which defines progressive disease (PD) as a 20% or more increase from measurement nadir, partial response (PR) as a decrease of 30% or more from baseline, and stable disease (SD) as the "middle ground." However, these classification criteria may not be optimal for any specific disease [91]. The aim of this study was to optimize the correlation between response classification and overall survival for MPM patients by varying the classification criteria.

The first step in this work was to quantify the relationship between response classification and survival for the standard RECIST classification criteria. We chose to use Harrell's *C* as our statistical metric instead of a log-rank survival metric because of the specific relationship trend we were trying to optimize. If response classification based on linear measurements were "perfectly" associated with survival, every patient classified as PR would live longer than every patient classified as SD, while both classes would live longer than every patient classified as PD. When this rela-

tionship holds true, $C = 1.0$. Using the modified RECIST measurement technique and the standard RECIST -30%/+20% criteria, we found a correlation of $C_{std} = 0.778$ between best patient response and survival.

While the performance of the standard RECIST criteria is within the range of "clinical utility" according to Harrell, performance could be improved by changing the response classification criteria to -64%/+50%. The performance of these criteria was measured as $C_{opt} = 0.855$. To avoid bias that may result from training and testing on the same group of patients, a cross-validation approach was used to estimate an unbiased performance of $C_{cv} = 0.829$, which is in the range of "high predictive accuracy." While comparing the full cohort performance (0.855) with the standard RECIST performance (0.778) yields a p-value of 0.039, the cross-validated p-value was 0.121. These p-values are calculated by considering the point estimates of C and their respective standard errors as well as the correlation between the two metrics. For a given group of patients, values of C from different classification criteria will be correlated because of the overlap between the discrete response categories. While C_{opt} is significantly larger than C_{std}, C_{cv} is still larger than C_{std}, though not significantly so. If the correlation between C_{std} and C_{cv} had not been taken into consideration, the p-value would have been 0.219 instead of 0.121.

Using the optimized response classification criteria, no patients are classified as having progressive disease as their best response. This is partly due to the broadening of the stable disease category; for a patient's best response to be in the realm of progressive disease with the standard RECIST classification criteria, the summed tumor thickness measurement needed to increase by at least 20%. With the proposed classification criteria, the summed tumor thickness measurement would need to increase by at least 50% to be considered progressive disease (a much higher hurdle than before).

While the lack of patients classified as PD is a byproduct of our particular patient sample, the effective reduction in response categories from three to two is actually in line with phase II trials, where classification into only two categories is common (e.g., "responders" and "non-responders").

In fact, if the optimization process above is conducted with only one cut-point to start instead of two, the same -64% criterion is obtained to separate a responders category from a non-responders category (see Figure 2.8). Some care also needs to be taken when interpreting the optimized criteria in terms of first follow-up response. With wider criteria, nearly all patients were classified as SD (77 of 78), since there has usually not been enough time for tumor burden to change dramatically in either direction. Despite the improved performance of the optimized criteria on the first follow-up response compared with standard RECIST, this study does not go so far as to advocate that all patient response should be assessed after only one follow-up scan and highlights that best response is usually achieved after a number of treatment cycles.

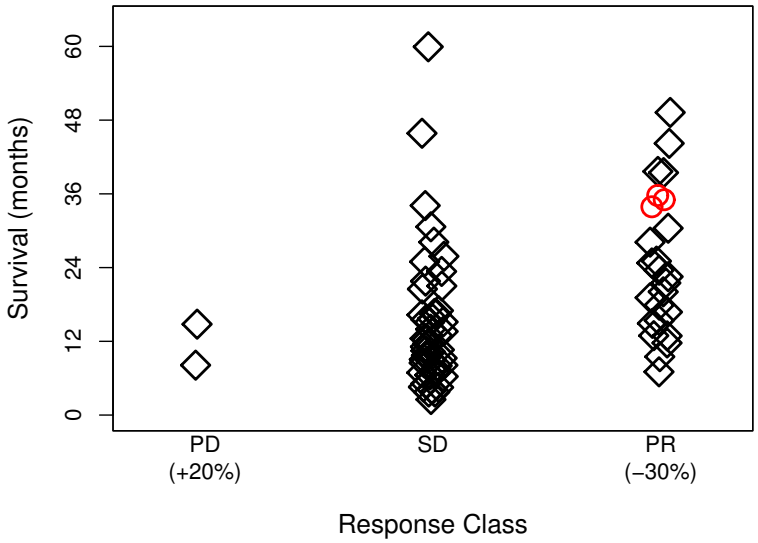
The issue of disease progression is also important in the context of initiating or withdrawing patient treatment. Some clinical trials incorporate progressive disease as an eligibility criterion, use progressive disease as a trigger to cease study treatment, and establish progression-free survival as an important endpoint. All these settings would be impacted by the classification criteria proposed in this study. Because this study was not originally an intervention study, we are unable to determine the impact of the proposed criteria on initiation of patient treatment. Of the 78 patients, 19 experienced disease progression according to the standard RECIST classification criteria at some point during their treatment, with a median time-to-progression of 5.0 months. Using the proposed classification criteria, however, only seven patients experienced disease progression at some point during their treatment with a median time-to-progression of 5.9 months. Using these revised criteria, patients may be eligible for clinical trials later and stay on treatment longer. In order to validate appropriate criteria for progression, it may be more appropriate to identify tumor thickness changes that correspond to meaningful deterioration in other patient-rated outcomes such as dyspnea, pain, and quality of life.

Previously, theoretical studies explored the possibility of alternate response criteria for MPM by investigating linear measurement cut-points in aspherical geometries [51]. Oxnard *et al.* obtained classification criteria of −67.9%/+100.1% for an annulus, −51.5%/+45.4% for a lens, and

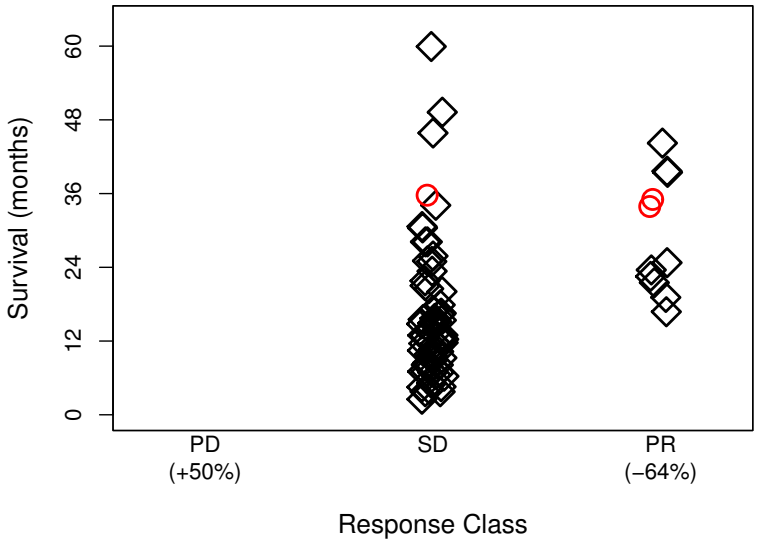
−65.8%/+73.6% for a crescent geometry, with linear measurements made according to the modified RECIST protocol. These alternate criteria are all substantially "wider" than the standard RECIST criteria, as are the optimized criteria we derived in this study; however, the theoretical criteria of Oxnard *et al.* were all based on volumetric equivalence to -30%/+20% changes in the diameter of a sphere, and the somewhat arbitrary provenance of those original criteria were outlined in section 1.2.1.

This study sought to identify classification criteria that optimized correlation with overall survival. To fully validate these new response criteria derived from this moderately sized database, they must be tested on a larger independent patient cohort. While the leave-one-out cross-validation used in this study attempts to simulate this process, it is not a substitute for a full independent validation, and future work will seek to validate these proposed response criteria. It should also be cautioned that while these criteria predict survival in patients on cytotoxic chemotherapy, it is unclear whether they would be a valid candidate surrogate for survival benefit in patients receiving a targeted therapy.

To summarize, the current standard for response assessment in patients with malignant pleural mesothelioma is a set of linear tumor thickness measurements acquired according to the modified RECIST protocol. Changes in these tumor measurements are compared with classification criteria, currently defined as -30% for partial response and +20% for progressive disease. Despite the original arbitrary provenance of these cut points, they perform adequately and are within the range of "clinical utility." However, by changing these criteria to -64% and +50%, respectively, the correlation between tumor response and overall survival is improved. These optimized classification criteria appear better suited to the specific morphology and growth pattern of mesothelioma and may prove useful in the assessment of clinical trials and routine patient care.



(a)



(b)

Figure 2.6: Correlation between best response classification per patient and survival. 2.6a, using the standard RECIST classification criteria. 2.6b, using the classification criteria derived from optimization on the full patient cohort. Black diamonds indicate observed events, while red circles indicate losses to follow-up. Performance metrics are summarized in Table 2.3.

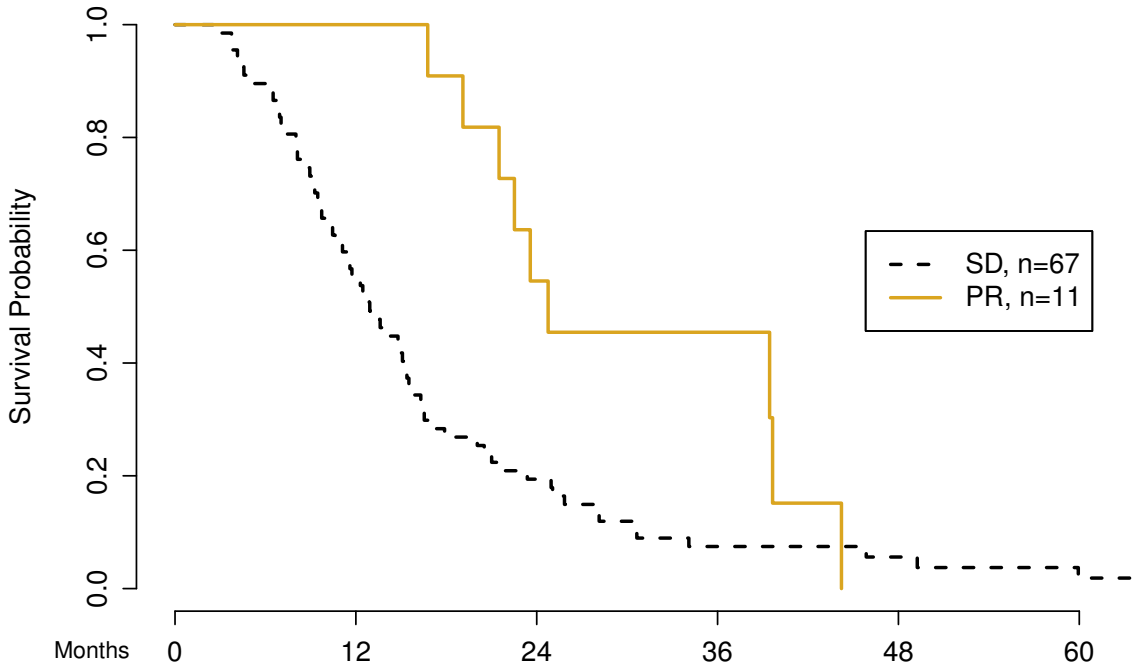


Figure 2.7: Overall survival from diagnosis by response category, using the optimized -64%/+50% response criteria and each patient's best response (stable disease, SD; partial response, PR). No patients were classified as having progressive disease with the optimized criteria.

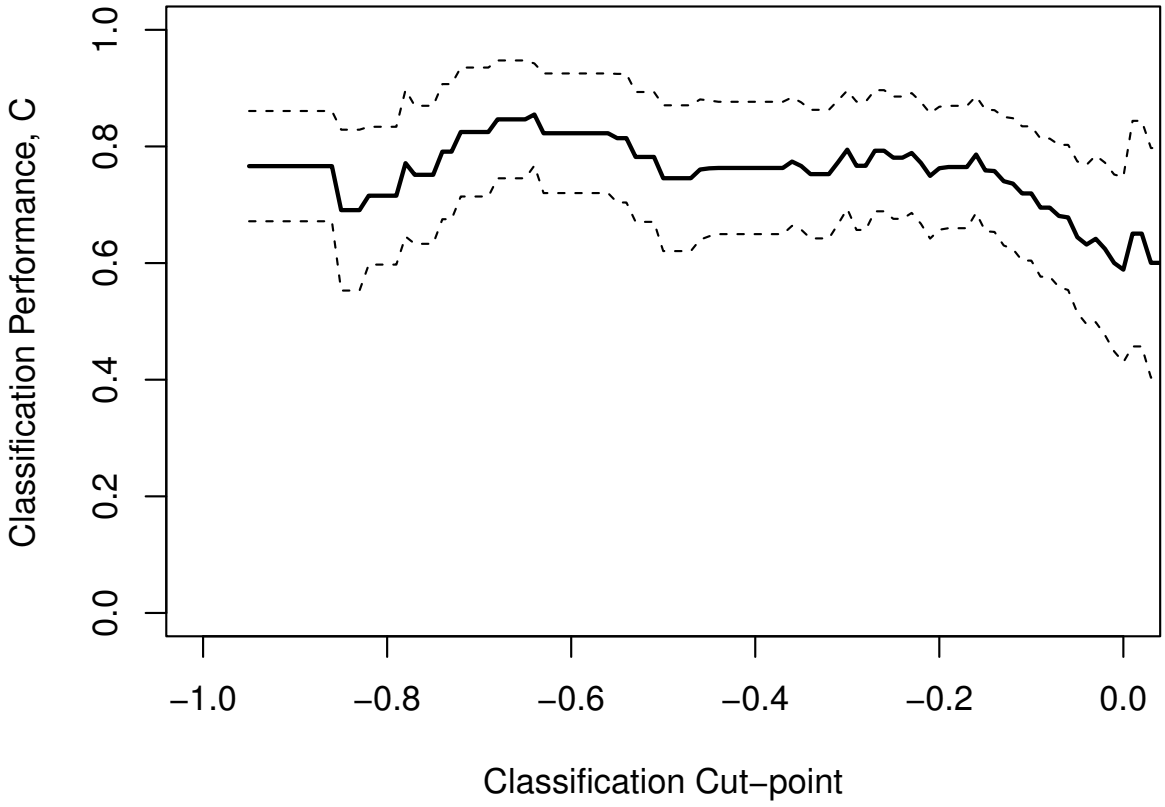


Figure 2.8: Classification performance (C) as a function of a single cut-point, where now response classification is *a priori* into only two groups ("responders" and "non-responders"). The solid line is the value of C for the cut-point, while the dashed lines are the point estimate 95% confidence intervals.

CHAPTER 3

AREA CONTOURS AS A POTENTIAL TOOL FOR RESPONSE ASSESSMENT

"Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease." – Sir William Osler

3.1 Introduction

The current clinical method for tumor response assessment in malignant pleural mesothelioma (MPM) is the modified Response Evaluation Criteria in Solid Tumors (RECIST) guidelines, which calls for two linear measurements of tumor thickness from each of three axial computed tomography (CT) sections to be summed as the tumor burden measurement [47]. It has been previously shown, however, that modified RECIST measurements have a relative inter-observer variability that can span a range of 30% under highly idealized image measurement conditions [52, 53]. This inter-observer variability is so large that from observer effects alone, a patient with truly stable disease (i.e., no actual change in tumor size between two time points) may be incorrectly classified as progressive disease or partial response (PD or PR, respectively). The level of variability is an especially salient point since clinically, multiple radiologists are often involved with the assessment of a given patient over time.

While the goal of the (modified) RECIST measurement system is to capture changes in overall tumor burden, and use these changes as a metric by which to gauge tumor response, the most complete measure of true tumor bulk is three-dimensional volume. Many tumor morphologies allow for a spherical approximation, where there is a clear mathematical relationship between a sphere's volume and radius. Therefore, for tumors such as lung nodules, the RECIST measurement system may be an appropriate tool to assess tumor bulk. For the highly non-spherical morphology

of MPM, though, one-dimensional measurements may not accurately capture changes in volume. Volume has been shown to be a significant predictor for overall and progression-free survival in patients with MPM [36], but the complete manual segmentation of MPM volumes on medical images is prohibitively time consuming for clinical implementation as a measurement tool. A study by Frauenfelder *et al.* [57] used a linear shape-based interpolation technique, requiring contours on "every fourth or fifth slice." The main conclusion of this study related to volumetric response was that the inter-observer agreement of volumetric response classification is much higher than for manual modified RECIST response classification (general $\kappa = 0.9$ vs general $\kappa = 0.33$, respectively). The comparison in that study was made using discretizations of changes in disease volume according to volumetric response criteria extrapolated from a spherical geometry, where the cutoffs for PR/PD were -65%/+73%. Therefore, the increased inter-observer agreement for volumetric measurements over linear measurements is likely in part due to the widening of the stable disease category alone (the standard RECIST criteria of -30%/+20% were used for linear measurements). The study by Liu *et al.* [58] utilized a combination of semi-automated techniques for volumetric MPM segmentation, and their analysis revealed changes in tumor volume to be significantly associated with patient survival.

From these studies, it is reasonable to believe that area (two-dimensional) measurements might serve as a middle ground; contouring visible disease on three axial CT sections (to replicate the three sections of modified RECIST) is less time consuming than full three-dimensional volume contours and should be a better representation of tumor bulk than one-dimensional measurements on those same three CT sections. Furthermore, the sum of three area measurements can be interpreted as a pseudo-volume for use in response assessment, and these pseudo-volumes may show the same improved inter-observer agreement exhibited by full volume measurements.

The goal of this study was to evaluate tumor area measurements on CT scans as a more complete and, potentially, less variable metric for response assessment for MPM. In this study, variability will be quantified for two distinct measurement tasks: baseline and follow-up measurements.

First, there is the variability in area (or pseudo-volume) measurement at a single time-point, i.e., baseline. When observers are presented with a "blank slate," their resulting contours will likely exhibit high variability. In the work by Armato *et al*. [52], linear tumor thickness measurements were all made originating from the same location between observers, and the variability was still fairly large. Due to the free-form nature of the baseline measurements, we expect to find a large variability in the single time-point measurements.

It may be more clinically relevant to measure the variability in follow-up measurements, since for a given patient, only one initial set of baseline measurements is made, and follow-up measurements are made with reference to the pre-existing measurements. It has been shown that initial outlines strongly influence inter-observer precision [92], and therefore the inter-observer variability will likely be reduced for follow-up measurements when compared with baseline measurements. Furthermore, changes in tumor measurements used for response assessment are often discretized by response classification criteria, and the inter-observer agreement in response classification will be quantified using relevant response criteria for area measurements.

3.2 Materials and Methods

3.2.1 Patient Cohort

The patient cohort in this study consisted of 31 patients with biopsy-proven MPM. The patients consisted of 27 males and 4 females, with a median age of 68 years at treatment initiation (range 49–81 years). All patients were part of a phase II clinical trial for a chemotherapy regimen consisting of cisplatin, pemetrexed, and bevacizumab (a monoclonal antibody that inhibits vascularization through the vascular endothelial growth factor, or VEGF, pathway) [93]. The HIPAA-compliant clinical trial was IRB-approved at our institution, and no scan was acquired for this area measurement study specifically. This patient cohort was chosen separately from the large cohort outlined in section 2.2.1 because the database of patients from Australia was not yet available when planning

for this study began.

3.2.2 Imaging

For each of the 31 patients, two scans were used in this study: first, the baseline scan acquired at most four weeks prior to treatment initiation, and second, the first follow-up scan acquired after two cycles of chemotherapy (median span between scans = 47 days). The diagnostic thoracic helical CT scans were performed on the Philips Brilliance 16-slice scanner (n=39), Brilliance 40-slice scanner (n=2), Brilliance 64-slice scanner (n=20), or Brilliance iCT (n=1) at our institution. Each CT section was reconstructed as a 512 x 512-pixel image matrix, with pixel dimensions ranging from 0.54–0.90 mm. Axially reconstructed slice thickness was three millimeters for all scans. For 28 of the patients, iodinated contrast media was used for both scans, while for one patient each contrast was used on the first scan only, on the second scan only, and on neither scan.

3.2.3 Area Measurement Acquisition

On each of the 31 baseline CT scans, three axial sections with visible disease were selected by one of the observers. Sections were selected with consideration for disease burden and proximity to anatomic landmarks for ease of follow-up measurements and were separated axially by at least one centimeter, in a similar manner to the process used to select the three axial sections used for the modified RECIST measurement technique. Using an in-house software package, Abras, five observers (all attending radiologists) independently contoured the visible disease on the 93 preselected baseline images. The radiologists were able to browse the entire CT image stack for each patient, as well as adjust window and level settings, but were only able to contour tumor on the preselected sections. Contours were converted to area measurements using Green's theorem [94], leading to 465 baseline area measurements. Before beginning the baseline area measurement process, all five observers were shown identical training materials for use of the software and were given identical written and verbal instructions for completing the study measurements. Observers

were advised to exclude regions of effusion and lung from their contours. For more information about Abras, see Appendix A of reference [95].

For the follow-up phase of the study, three of the original five observers were presented with the baseline scan for each patient and were able to see the baseline contours on the three previously selected axial sections. All three observers were presented with the same set of baseline contours for each patient (all baseline contours were taken from the observer who had initially selected the three axial sections to contour at baseline). The follow-up observers could not alter these baseline contours but could see the entire baseline scan for each patient as reference. The follow-up measurement process consisted of two steps: first, each observer needed to independently find the matched axial follow-up section for each of the three baseline scan sections on which measurements had been made, and second, the observer needed to construct the follow-up contours to capture tumor area on the selected follow-up sections. This process replicates the clinical workflow typically used for response assessment, where the observer tasked with making follow-up measurements is able to visualize the previous measurements, but is wholly responsible for the placement of the new measurements. Again, contours were converted to enclosed area measurements.

For both the baseline and follow-up time-points for each patient, area measurements were analyzed both as individual sections and as the sum of section measurements per patient, which is more clinically relevant as a representation of tumor bulk (pseudo-volume). In total, there were 744 individual area measurements (465 baseline and 279 follow-up measurements), or 248 pseudo-volume measurements (155 baseline and 93 follow-up measurements).

3.2.4 Data Analysis

3.2.4.1 Baseline Measurement Analysis

Estimating the variation in area measurements attributable to differences between observers was accomplished using a random effects analysis of variance (ANOVA) model [96]. We start with a

linear model for the section-by-section data,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{k(j)} + \varepsilon_{ijk}, \tag{3.1}$$

where y_{ijk} represents the measurement for the i^{th} observer (i = 1–5) in the j^{th} patient (j = 1–31) on the k^{th} section (k = 1–3). In equation 3.1, μ represents the overall mean, α_i represents the effects of the observers, β_j represents the effects of the different patients, $\gamma_{k(j)}$ represents the effects of the different sections nested within each patient (i.e., subscripts k are only meaningful for a specified subscript j), and ε_{ijk} are the residual errors. For summed-area measurements (the sum of three sections per patient, representing a composite tumor area), the linear model for measurements is

$$z_{ij} = m + a_i + b_j + e_{ij}, \tag{3.2}$$

where the explanation of terms remain the same as in equation 3.1 but variables have changed from Greek to Roman to avoid any confusion (z_{ij} is the summed measurement, m is the overall mean, a_i is the observer effect, b_j is the patient effect, and e_{ij} is the residual error). Note the absence of a section-effect term in equation 3.2.

Since μ (or m) is the overall mean, the remaining effects are assumed to be normally distributed with zero mean, and the variance component attributable to each effect is estimated in the ANOVA model. All analyses were performed using the academic edition of Revolution R Enterprise (version 4.3, based on R version 2.12) [86]. Once we obtained estimates for the variance components that involve the observer, $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\varepsilon$ ($\hat{\sigma}^2_a$ and $\hat{\sigma}^2_e$ for summed-area variance estimates), the absolute inter-observer variability is calculated from standard rules of probability. Following a similar derivation to the work by Armato *et al.* [52], absolute inter-observer variability for section-by-section measurements (i.e., variance in the difference of per-section area measurements between

any two observers) is given by

$$\begin{aligned} \mathrm{Var}\left[y_{i'jk} - y_{ijk}\right] &= \mathrm{Var}\left[(\alpha_{i'} - \alpha_i) + \left(\varepsilon_{i'jk} - \varepsilon_{ijk}\right)\right] \\ &= 2\left(\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2\right), \end{aligned} \tag{3.3}$$

and a similar equation can be derived for the summed measurement data. Once this absolute inter-observer variability is derived, which we will henceforth call $\hat{\sigma}_y^2$ or $\hat{\sigma}_z^2$ for the section-by-section and summed measurement data, respectively, the 95% confidence interval on absolute inter-observer variability can be calculated as $\pm 1.96\,\hat{\sigma}_y$.

For relative inter-observer variability, we are interested in the variability of the quantity

$$\frac{\left(y_{i'jk} - y_{ijk}\right)}{y_{ijk}} = \frac{y_{i'jk}}{y_{ijk}} - 1.$$

Finding the variability of this quantity requires estimating the variance of the ratio $y_{i'jk}/y_{ijk}$, which can be accomplished with a log transformation using

$$\begin{aligned} \mathrm{Var}\left[\ln\left(\frac{y_{i'jk}}{y_{ijk}}\right)\right] &= \mathrm{Var}\left[\ln\left(y_{i'jk}\right) - \ln\left(y_{ijk}\right)\right] \\ &= 2\left(\hat{\sigma}_\alpha'^2 + \hat{\sigma}_\varepsilon'^2\right), \end{aligned} \tag{3.4}$$

where the $\hat{\sigma}_\star'^2$ are derived from fitting equation 3.1 with $\ln\left(y_{ijk}\right)$ instead of y_{ijk}. If we denote the result from equation 3.4 as $\hat{\sigma}_y'^2$, we can calculate the 95% confidence interval on $\ln\left(y_{i'jk}/y_{ijk}\right)$ as $\pm 1.96\,\hat{\sigma}_y'$. Finally, inverting the log transformation, the 95% confidence interval on the relative inter-observer variability is

$$e^{-1.96\,\hat{\sigma}_y'} - 1 < \frac{\left(y_{i'jk} - y_{ijk}\right)}{y_{ijk}} < e^{+1.96\,\hat{\sigma}_y'} - 1. \tag{3.5}$$

A similar derivation can be followed for summed measurement data.

From the same linear models fit using equations 3.1 and 3.2, the full set of estimated variance components are produced. These variance components are used to construct the intra-class correlation (ICC), which can be interpreted as a generalized R^2, or proportion of total variation attributable to a specified source [97,98]. For instance, the proportion of total variation attributable to patient effects in the summed area measurement model would be

$$ICC_{pat}^{sum} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + \hat{\sigma}_e^2}. \tag{3.6}$$

If ICC_{pat}^{sum} is equal to unity, all the variability in the data can be attributed to differences between patients, and none is attributable to differences between observers. Similarly, the proportion of total variability attributable to reader effects in the per-section area measurement model would be

$$ICC_{obs}^{slice} = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\beta^2 + \hat{\sigma}_\gamma^2 + \hat{\sigma}_\varepsilon^2}. \tag{3.7}$$

Finally, the agreement between observers can also be quantified using the mean value of the Spearman rank correlation statistic (ρ) between each pair of readers.

3.2.4.2 Follow-Up Measurement Analysis

The follow-up area measurements can be analyzed with similar methodology to the baseline measurements. The same linear models and intra-class correlations will be calculated for both the per-section area measurements and summed area measurements. The main expected difference between baseline and follow-up measurements is the "biasing" effect of seeing previous contours. Since all observers will be drawing follow-up contours while seeing the same set of baseline contours, it can be expected that the amount of inter-observer variability will be reduced at follow-up.

The inter-observer variabilities (especially the relative variability numbers derived from equation 3.5 for follow-up measurements) will be best understood in the context of response classification criteria for area measurements. The 1981 World Health Organization (WHO) criteria

for bi-dimensional measurement changes called for two linear measurements to be made (longest diameter and longest perpendicular diameter), and their product represented the bi-dimensional measurement of interest [42]. The response criteria for WHO measurements were given as a measurement decrease of 50% or more for PR, a measurement increase of 25% or more for PD, and SD for a measurement change between -50% and +25%.

Alternate bi-dimensional measurement classification criteria are given in Oxnard *et al.* [51] for geometric models beyond the sphere, shown in Figure 3.1. The -50%/+25% WHO criteria become -81.2%/+163.0%, -65.3%/+71.0%, and -79.3%/+119.6% for the annulus, lens, and crescent, respectively. While these criteria are for bi-dimensional measurements (product of two linear measurements) and not true area measurements, the criteria will be used to determine if the derived relative inter-observer variabilities are so large as to lead to patients being misclassified due to observer variation alone.

The response classification criteria can also be used to rate the agreement between observers using Fleiss' Kappa statistic, which quantifies agreement between a fixed number of raters using categorical ratings [99]. Once response classification has been calculated for each patient for each observer, κ quantifies to what extent the observers agree with one another in terms of classification (and not in terms of actual area measurements). Specifically, κ is a measure of the agreement between the fixed observers that exceeds the agreement expected due to chance alone. If there is perfect agreement between observers in terms of classification, $\kappa = 1$, while $\kappa \leq 0$ if the agreement between observers is less than or equal to the agreement due to chance alone. There do exist some numerical guidelines for the interpretation of κ values [100], but these guidelines are completely arbitrary, and modern recommendations are that they should be avoided [101].

3.3 Results

For the 31 patients in this study, the median survival as assessed from the initiation of treatment was 12.5 months, with a 95% confidence interval of 9.3–16.8 months by Kaplan-Meier analysis.

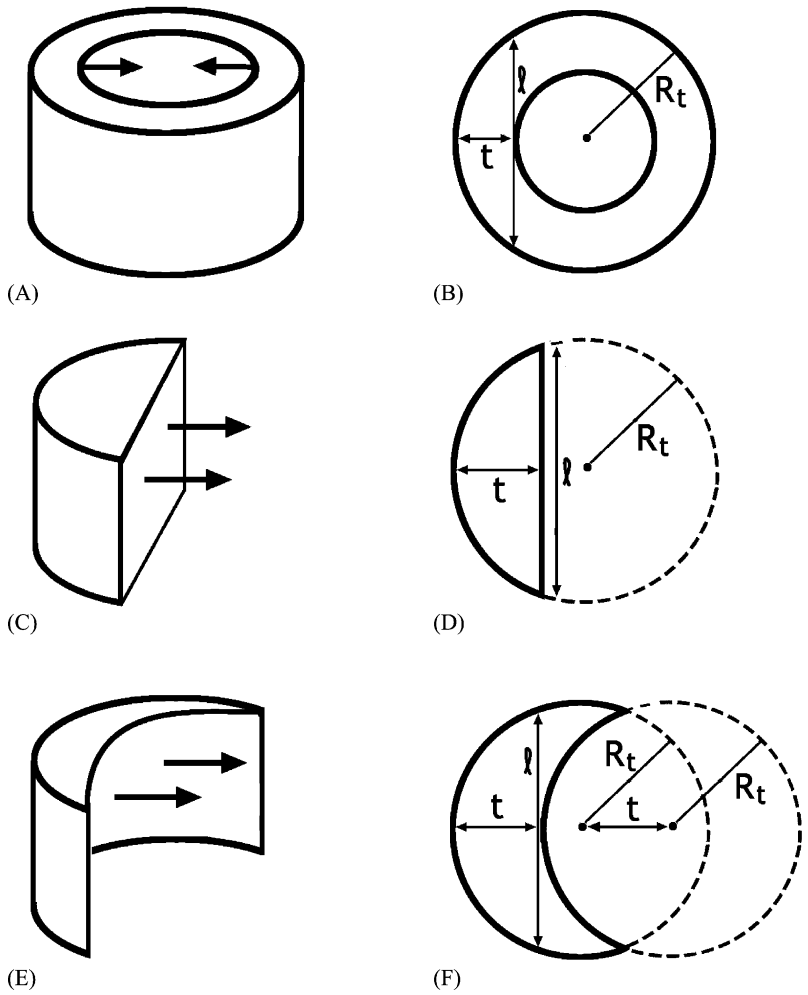


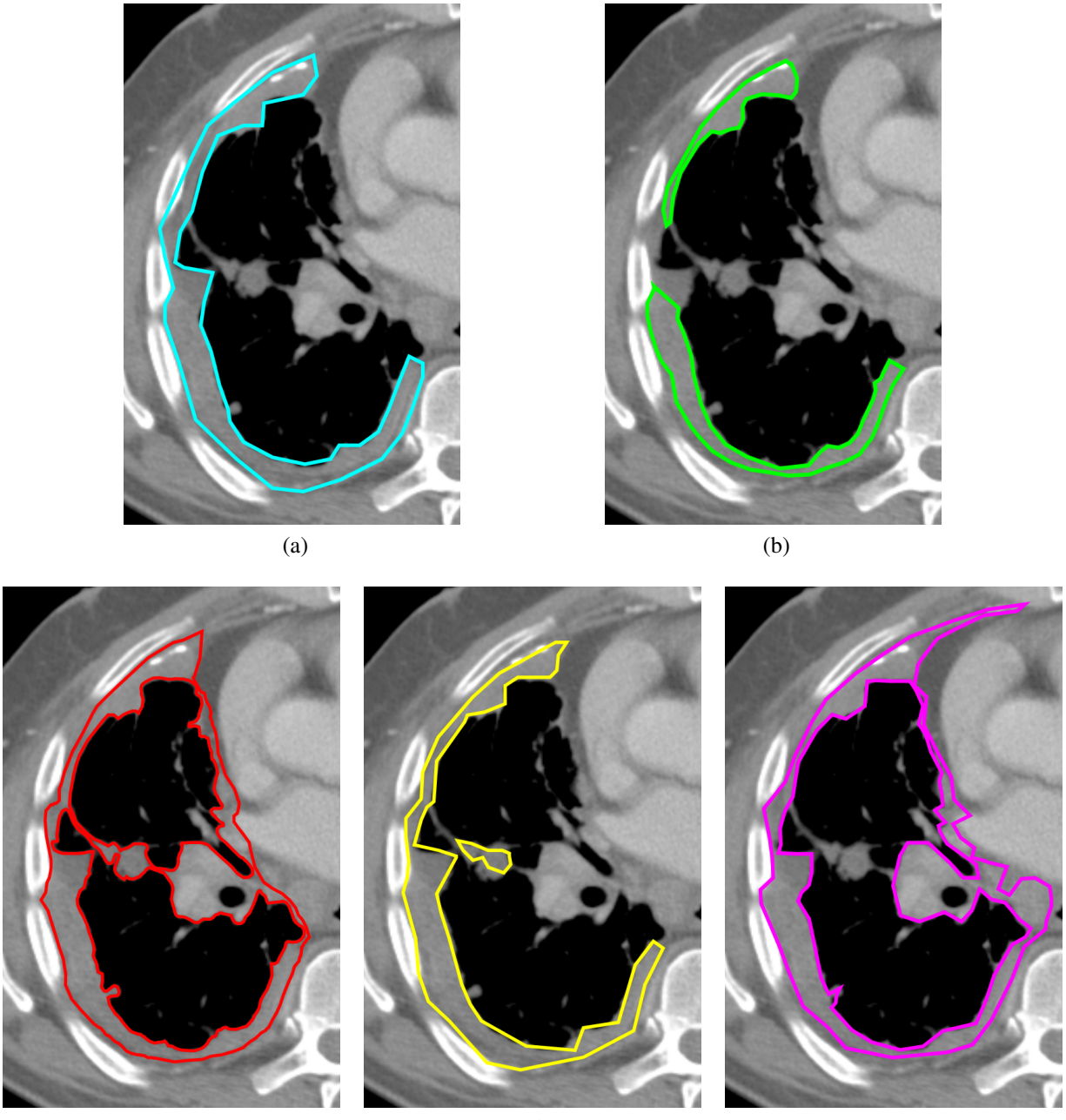
Figure 3.1: Three tumor models developed to represent mesothelioma-specific tumor geometry and growth, from Oxnard *et al*. [51]: (A and B), the annulus, (C and D), the lens, and (E and F), the crescent. (A, C, and E) Arrows indicate the direction of tumor growth. (B, D, and F) Cross-sectional views of each tumor model, where the product of tumor thickness t and tumor length l forms the bi-dimensional measurement of interest in this study. Figure reprinted with permission of the publisher.

Twenty-eight out of 31 deaths were observed, while for the remaining three patients the median follow-up was 31.1 months.

3.3.1 Baseline Measurements

The mean baseline per-section area measurement across five observers and 31 patients (with three individual area measurements per patient) was 2562 mm^2, and the mean summed area measurement across five observers and 31 patients (with one summed area measurement per patient) was 7686 mm^2. The baseline area contours of the five observers on the same CT section are shown in Figure 3.2. Figure 3.3 shows a plot of baseline summed area measurement data that depicts the difference between an individual observer's measurement and the average of all five observers' summed area measurements for a given patient versus the group average. This type of plot is similar to a Bland-Altman plot, but it should be noted that it is not identical [102, 103]. The mean difference between the individual observer's summed area measurements and the mean of the *other* four observers' summed area measurements is 678 mm^2, -2510 mm^2, 1885 mm^2, -1246 mm^2, and 1194 mm^2 for observer 1, 2, 3, 4, and 5, respectively. This can be easily seen in Figure 3.3, where, for example, observer 2 is clearly biased lower than the other observers. The average bivariate rank correlation between observers (i.e., average value of ρ for each pairwise comparison) for baseline summed area measurements was $\bar{\rho}^{sum} = 0.898$ ($p < 0.0001$).

Figure 3.4 shows a similar plot for baseline per-section area measurement data. The mean difference between the individual observer's per-section area measurements and the mean of the other four observers' per-section area measurements is 226 mm^2, -837 mm^2, 628 mm^2, -415 mm^2, and 398 mm^2 for observer 1, 2, 3, 4, and 5, respectively. The average bivariate rank correlation between observers for baseline per-section area measurements was $\bar{\rho}^{slice} = 0.885$ ($p < 0.0001$).



(a) (b) (c) (d) (e)

Figure 3.2: Five observers' outlines of malignant pleural mesothelioma on a single baseline CT section. The corresponding areas of these five measurements are (3.2a) 2756 mm^2, (3.2b) 1583 mm^2, (3.2c) 3877 mm^2, (3.2d) 2545 mm^2, and (3.2e) 3838 mm^2.

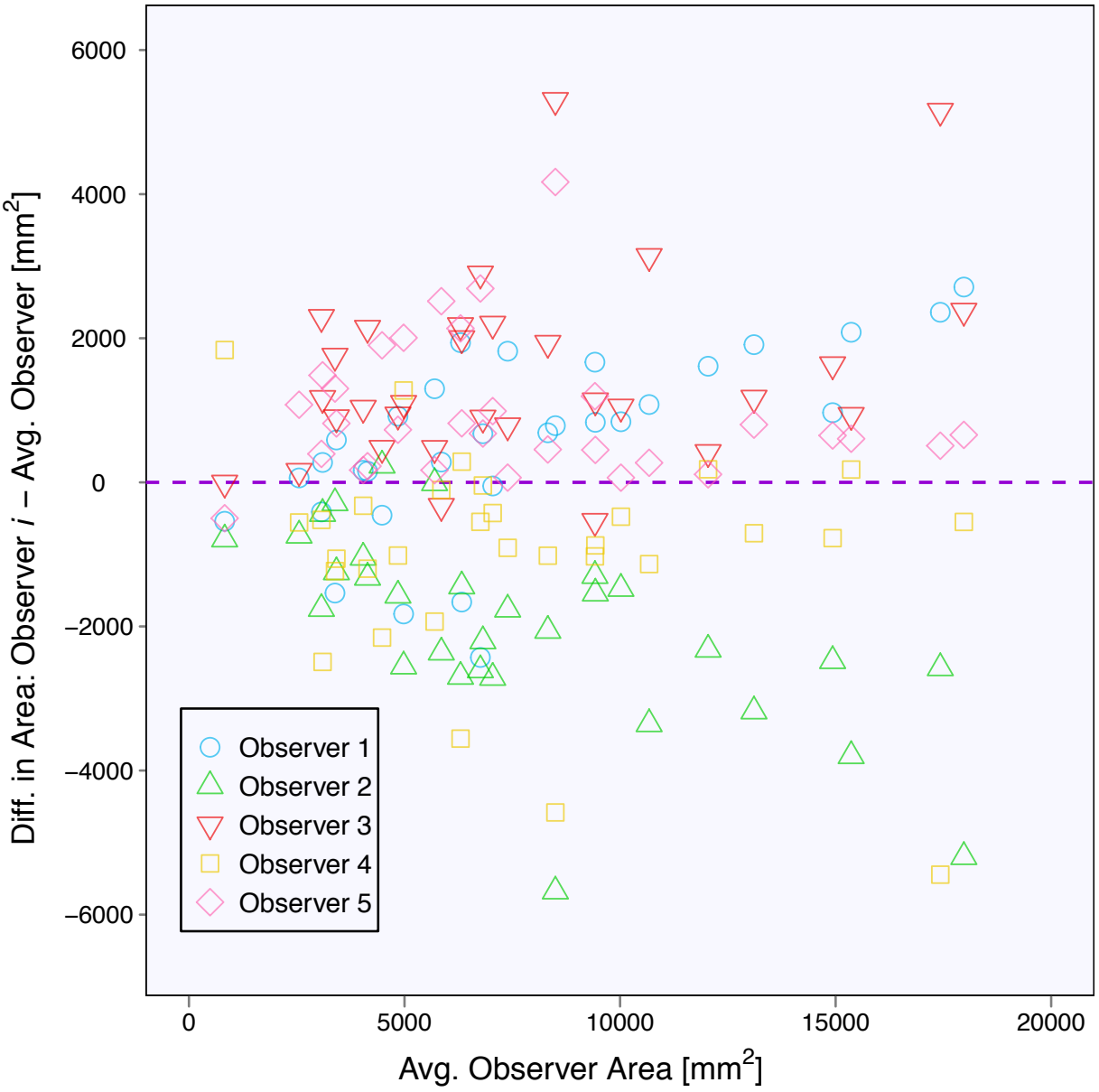


Figure 3.3: Plot of baseline summed area measurement data for 31 patients and five observers. The y-axis is the measurement difference between a given observer and the average of all observers, and the x-axis is the average of all observers.

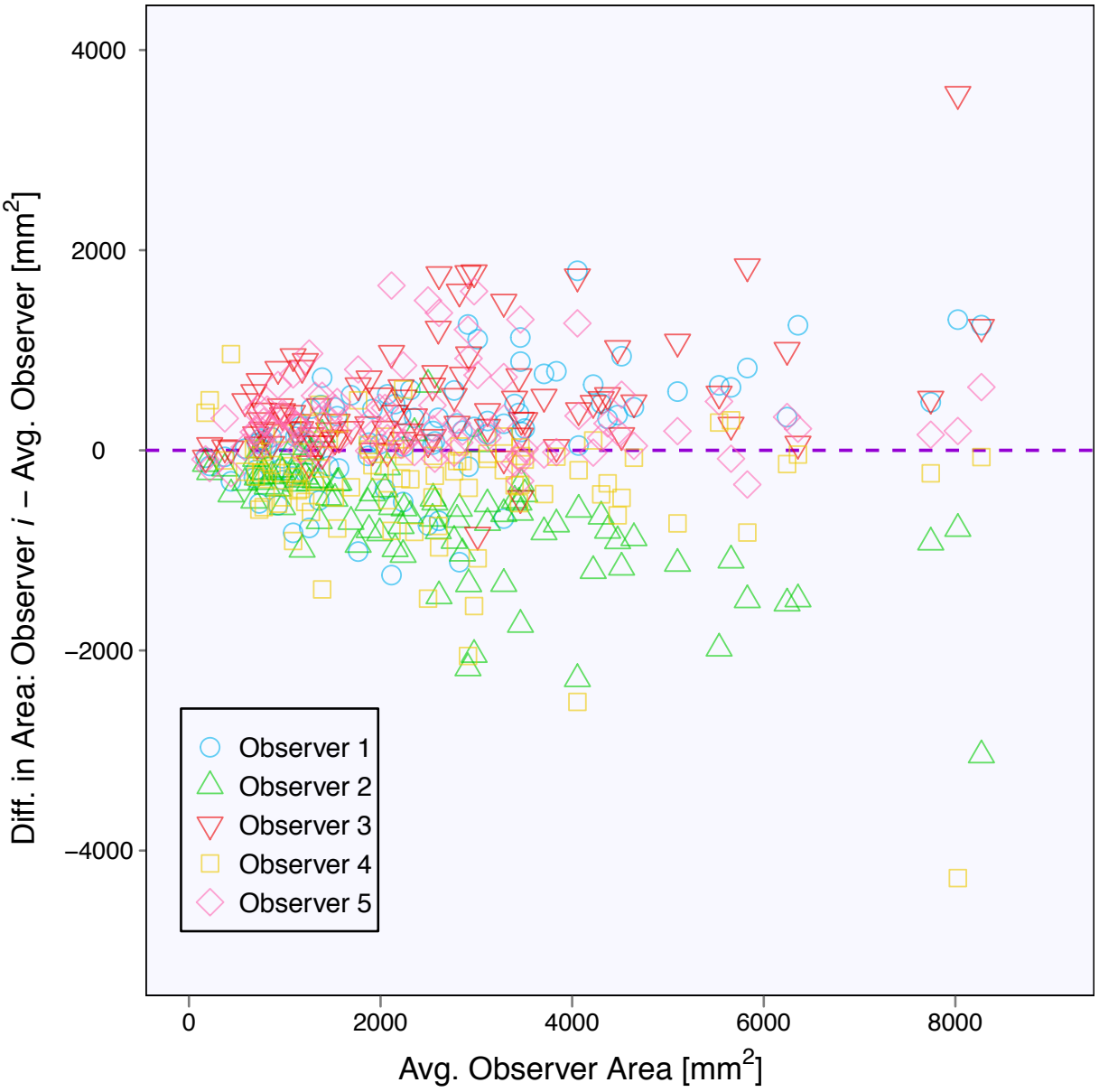


Figure 3.4: Plot of baseline per-section area measurement data for 31 patients, three slices per patient, and five observers. The y-axis is the measurement difference between a given observer and the average of all observers, and the x-axis is the average of all observers.

Results from the random effects model give the following per-section area measurement variance component estimates. For absolute variability, $\hat{\sigma}_\alpha = 482\,\text{mm}^2$ and $\hat{\sigma}_\varepsilon = 631\,\text{mm}^2$, yielding $\hat{\sigma}_y = 1122\,\text{mm}^2$. Therefore, the 95% confidence interval for absolute inter-observer variability is $\left[-2200\,\text{mm}^2, 2200\,\text{mm}^2\right]$, or $\pm 86\%$ of the mean per-section area measurement value.

The summed area measurement variance component estimates, for absolute variability, are $\hat{\sigma}_a = 1435\,\text{mm}^2$, $\hat{\sigma}_e = 1436\,\text{mm}^2$, yielding $\hat{\sigma}_z = 2871\,\text{mm}^2$. Therefore, the 95% confidence interval for absolute inter-observer variability is $\left[-5627\,\text{mm}^2, 5627\,\text{mm}^2\right]$ for the summed baseline area measurements, or $\pm 73\%$ of the mean summed area measurement value. These values are summarized in Tables 3.1 and 3.3.

The relative variance component estimates for the baseline per-section area measurements are $\hat{\sigma}'_\alpha = 0.230$ and $\hat{\sigma}'_\varepsilon = 0.367$, giving $\hat{\sigma}'_y = 0.612$. This gives a 95% confidence interval for relative inter-observer variability of [-70%, +232%] from equation 3.5. For the relative baseline summed area measurement variance component estimates, we have $\hat{\sigma}'_a = 0.242$ and $\hat{\sigma}'_e = 0.339$, giving $\hat{\sigma}'_z = 0.624$. This gives a 95% confidence interval for relative inter-observer variability of [-71%, +240%] (i.e., one would expect to find, with 95% confidence, that a given observer's summed area measurement for a given patient would be up to 71% less than or 240% greater than another observer's summed area measurement for the same patient). These results are summarized in Tables 3.2 and 3.3.

Variance Component		SD [mm^2]	Variance [mm^4]
Per-Section Measurements			
	Observer, α	481.7	232017
	Patient, β	1327.5	1762245
	Slice, γ	1135.5	1289430
	Residual, ε	630.8	397888
	$\hat{\sigma}_y^2 = 2\left(\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2\right)$	1122.4	1259810
Summed Measurements			
	Observer, a	1435.3	2060216
	Patient, b	4422.1	19554923
	Residual, e	1435.7	2061179
	$\hat{\sigma}_z^2 = 2\left(\hat{\sigma}_a^2 + \hat{\sigma}_e^2\right)$	2871.0	8242790

Table 3.1: Estimated variance components for the absolute baseline area measurement linear models, given both as standard deviations (SD) and variances.

Variance Component		SD	Variance
Per-Section Measurements			
	Observer, α'	0.230	0.053
	Patient, β'	0.681	0.463
	Slice, γ'	0.452	0.204
	Residual, ε'	0.367	0.135
	$\hat{\sigma}_y'^2 = 2\left(\hat{\sigma}_\alpha'^2 + \hat{\sigma}_\varepsilon'^2\right)$	0.612	0.375
Summed Measurements			
	Observer, a'	0.242	0.059
	Patient, b'	0.741	0.549
	Residual, e'	0.369	0.136
	$\hat{\sigma}_z'^2 = 2\left(\hat{\sigma}_a'^2 + \hat{\sigma}_e'^2\right)$	0.624	0.389

Table 3.2: Estimated variance components for the relative baseline area measurement linear models, given both as standard deviations (SD) and variances.

95% Confidence Interval	Lower Bound	Upper Bound	Range
Absolute, Per-Section	-2200 mm^2	2200 mm^2	4400 mm^2
Absolute, Summed	-5627 mm^2	5627 mm^2	11254 mm^2
Absolute (as % of mean), Per-Section	-86%	86%	172%
Absolute (as % of mean), Summed	-73%	73%	146%
Relative, Per-Section	-70%	232%	302%
Relative, Summed	-71%	240%	311%

Table 3.3: 95% confidence intervals (CI) for absolute and relative inter-observer variabilities for baseline area measurements of malignant pleural mesothelioma. CI for both per-section measurements and summed (3 per patient) measurements are assessed using the methods in section 3.2.4.

Intra-class correlation statistics were also computed for the baseline area measurement data using the estimated variance components in Table 3.1. These values are summarized in Table 3.4. For the per-section data, the inter-patient variability accounts for 82.7% of the total variability, while the inter-observer variability accounts for only 6.4% of the total variability. For the summed data, the inter-patient variability accounts for 82.6% of the total variability, while the inter-observer variability accounts for only 8.7% of the total variability. For both the per-section and summed data, the inter-patient variability is the significant majority of total variability as expected from the inherently wide range of disease sizes between patients. While the inter-observer variability appears to constitute a small fraction of overall variability for per-section and summed data in Table 3.4, both inter-observer variabilities are significantly larger than zero, and the total variability

ICC Statistic	Value	95% Confidence Interval
ICC_{pat}^{slice}	0.872	[0.710, 0.894]
ICC_{obs}^{slice}	0.064	[0.022, 0.365]
ICC_{pat}^{sum}	0.826	[0.635, 0.917]
ICC_{obs}^{sum}	0.087	[0.027, 0.452]

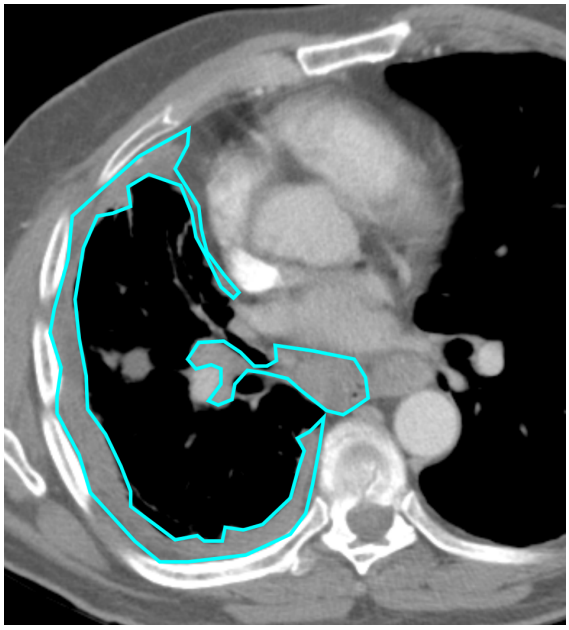
Table 3.4: Intra-class correlation statistics for the baseline area measurement data, calculated from Table 3.1.

including both inter-observer and inter-patient effects is substantial since the range of disease sizes in the patient population is quite wide.

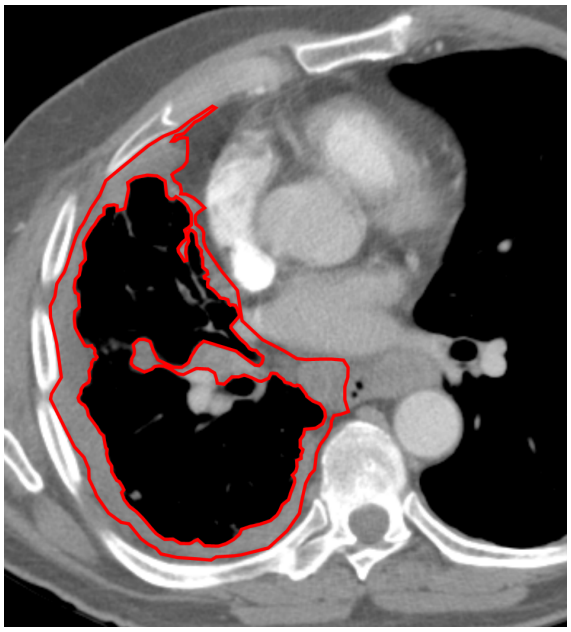
3.3.2 Follow-up Measurements

The mean follow-up per-section area measurement across observers and patients was 2532 mm^2, and the mean summed area measurement was 7597 mm^2. On average, the follow-up per-section area measurements were 7.5% lower than the baseline areas. Similarly, follow-up summed area measurements were 9.7% lower than the corresponding baseline measurements on average. Follow-up contours are shown in Figure 3.5 for the same baseline axial section shown in Figure 3.2 (specifically, Figure 3.2e was the baseline contour shown to follow-up observers for this patient case).

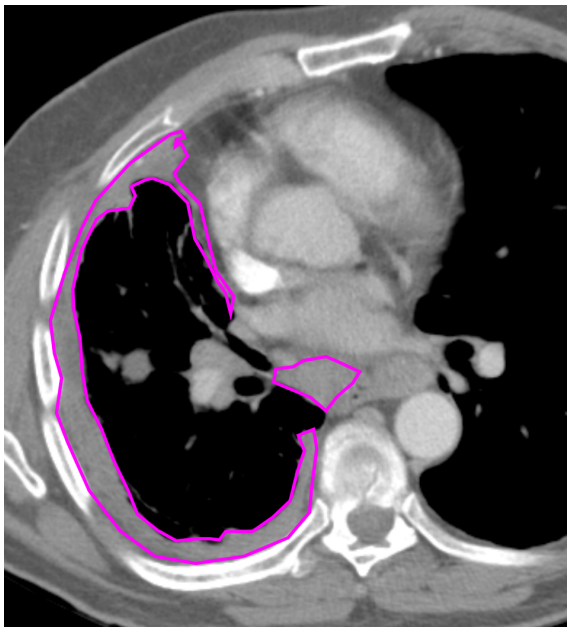
Since the follow-up observers were free to choose the axial section for a given contour, there was some disagreement in axial slice selection. On average, two different slices were contoured at follow-up for a given baseline contour across the three observers; for 21 of the 93 baseline contours, all three observers selected a different axial slice for follow-up measurement. The mean deviation in slice location between observers was 0.72 slices, or approximately 2.2 mm. Figure 3.6



(a)



(b)



(c)

Figure 3.5: Three observers' follow-up outlines of malignant pleural mesothelioma on a single CT section, matched to the section shown in Figure 3.2 (all observers were shown the baseline contour from Figure 3.2e). The corresponding areas of the three measurements are (3.5a) 3813 mm^2, (3.5b) 3659 mm^2, and (3.5c) 3139 mm^2.

shows a plot of follow-up summed area measurement data that depicts the difference between an individual observer's measurement and the average of all three observers' summed area measurements for a given patient versus the group average. The mean difference between the individual observer's summed area measurements and the mean of the *other* two observers' summed area measurements is -430 mm^2, 33 mm^2, and 397 mm^2 for observer 1, 2, and 3, respectively. Figure 3.7 shows a similar plot for follow-up per-section area measurement data. The mean difference between the individual observer's per-section area measurements and the mean of the other two observers' per-section area measurements is -143 mm^2, 11 mm^2, and 132 mm^2 for observer 1, 2, and 3, respectively.

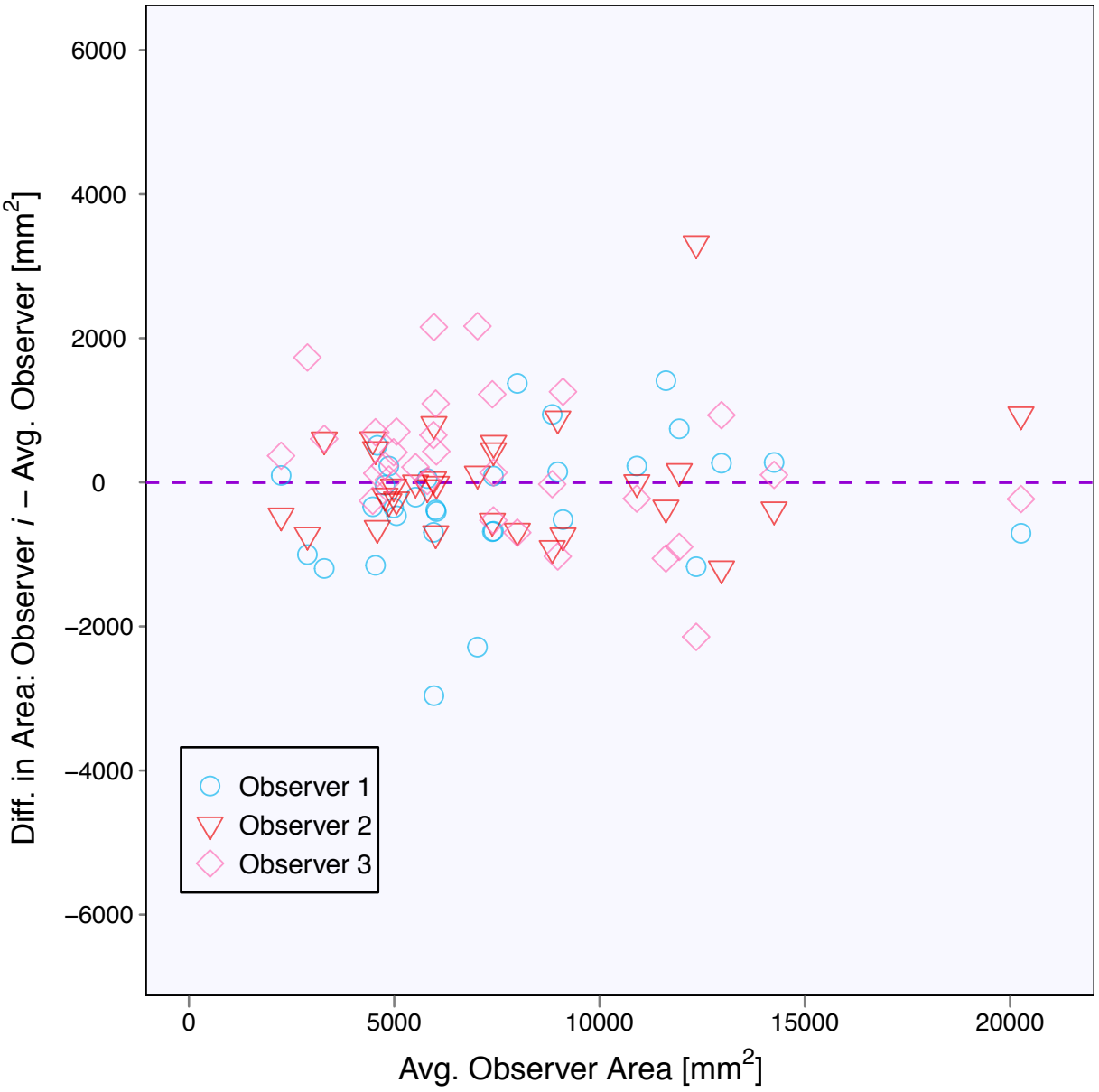


Figure 3.6: Plot of follow-up summed area measurement data for 31 patients and 3 observers. The y-axis is the measurement difference between a given observer and the average of all observers, and the x-axis is the average of all observers. Note that the y-axis is identical to Figure 3.3, highlighting the increased agreement among observers for follow-up measurements.

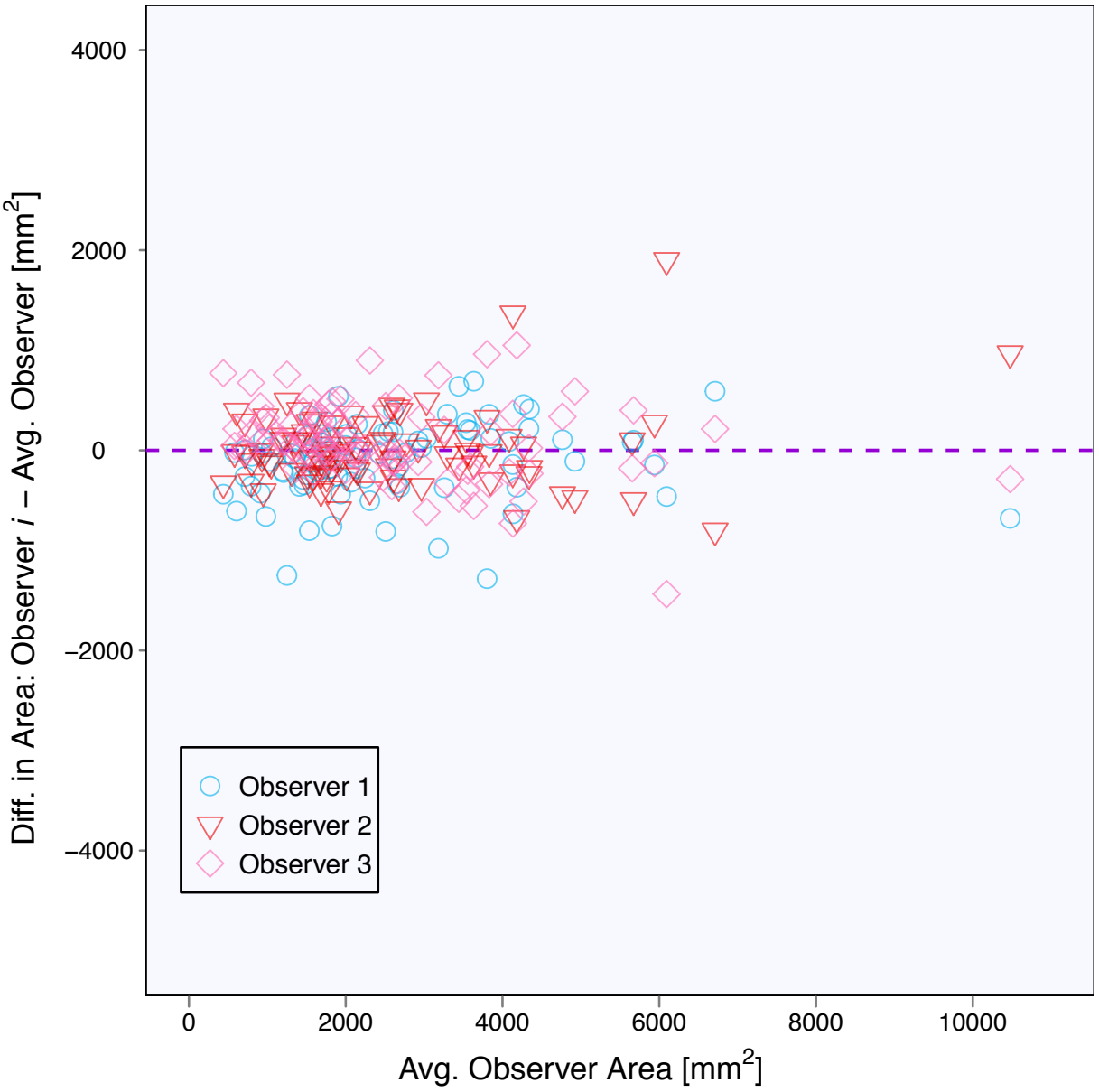


Figure 3.7: Plot of follow-up per-section area measurement data for 31 patients, 3 slices per patient, and 3 observers. The y-axis is the measurement difference between a given observer and the average of all observers, and the x-axis is the average of all observers. Note that the y-axis is identical to Figure 3.4, highlighting the increased agreement among observers for follow-up measurements.

Results from the random effects model give the following per-section area measurement variance component estimates. For absolute variability, $\hat{\sigma}_\alpha = 79\,\text{mm}^2$ and $\hat{\sigma}_\varepsilon = 461\,\text{mm}^2$, yielding

$\hat{\sigma}_y = 661\,\text{mm}^2$. Therefore, the 95% confidence interval for absolute inter-observer variability is $\left[-1296\,\text{mm}^2, 1296\,\text{mm}^2\right]$, or $\pm 51\%$ of the mean per-section area measurement value. The summed area measurement variance component estimates, for absolute variability, are $\hat{\sigma}_a = 194\,\text{mm}^2$, $\hat{\sigma}_e = 1096\,\text{mm}^2$, yielding $\hat{\sigma}_z = 1574\,\text{mm}^2$. Therefore, the 95% confidence interval for absolute inter-observer variability is $\left[-3085\,\text{mm}^2, 3085\,\text{mm}^2\right]$ for the summed follow-up area measurements, or $\pm 41\%$ of the mean summed area measurement value. These values are summarized in Tables 3.5 and 3.7.

The relative variance component estimates for the follow-up per-section area measurements are $\hat{\sigma}'_\alpha = 0.063$ and $\hat{\sigma}'_\varepsilon = 0.242$, giving $\hat{\sigma}'_y = 0.354$. This gives a 95% confidence interval for relative inter-observer variability of [-50%, +100%] from equation 3.5. For the relative follow-up summed area measurement variance component estimates, we have $\hat{\sigma}'_a = 0.061$ and $\hat{\sigma}'_e = 0.179$, giving $\hat{\sigma}'_z = 0.268$. This gives a 95% confidence interval for relative inter-observer variability of [-41%, +69%]. These results are summarized in Tables 3.6 and 3.7.

Variance Component		SD [mm^2]	Variance [mm^4]
Per-Section Measurements			
	Observer, α	78.7	6195
	Patient, β	1103.1	1216811
	Slice, γ	1143.7	1308134
	Residual, ε	460.9	212472
	$\hat{\sigma}_y^2 = 2\left(\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2\right)$	661.3	437336
Summed Measurements			
	Observer, a	193.8	37567
	Patient, b	3832.5	14687977
	Residual, e	1096.0	1201228
	$\hat{\sigma}_z^2 = 2\left(\hat{\sigma}_a^2 + \hat{\sigma}_e^2\right)$	1574.0	2477590

Table 3.5: Estimated variance components for the absolute follow-up area measurement linear models, given both as standard deviations (SD) and variances.

Variance Component		SD	Variance
Per-Section Measurements			
	Observer, α'	0.063	0.004
	Patient, β'	0.433	0.187
	Slice, γ'	0.405	0.164
	Residual, ε'	0.242	0.059
	$\hat{\sigma}_y'^2 = 2\left(\hat{\sigma}_\alpha'^2 + \hat{\sigma}_\varepsilon'^2\right)$	0.354	0.125
Summed Measurements			
	Observer, a'	0.061	0.004
	Patient, b'	0.486	0.236
	Residual, e'	0.179	0.032
	$\hat{\sigma}_z'^2 = 2\left(\hat{\sigma}_a'^2 + \hat{\sigma}_e'^2\right)$	0.268	0.072

Table 3.6: Estimated variance components for the relative follow-up area measurement linear models, given both as standard deviations (SD) and variances.

95% Confidence Interval	Lower Bound	Upper Bound	Range
Absolute, Per-Section	-1296 mm^2	1296 mm^2	2592 mm^2
Absolute, Summed	-3085 mm^2	3085 mm^2	6170 mm^2
Absolute (as % of mean), Per-Section	-51%	51%	102%
Absolute (as % of mean), Summed	-41%	41%	82%
Relative, Per-Section	-50%	100%	150%
Relative, Summed	-41%	69%	110%

Table 3.7: 95% confidence intervals (CI) for absolute and relative inter-observer variabilities for follow-up area measurements of malignant pleural mesothelioma. CI for both per-section measurements and summed (3 per patient) measurements are assessed using the methods in section 3.2.4.

The values of the ICC statistics for follow-up measurements are summarized in Table 3.8. For the per-section data, the inter-patient variability accounts for 92% of the total variability, while the inter-observer variability accounts for only 0.2% of the total variability. For the summed data, the inter-patient variability also accounts for 92% of the total variability, while the inter-observer variability accounts for only 0.2% of the total variability. For both the per-section and summed data, the inter-patient variability is the significant majority of total variability. For these follow-up data, the inter-observer variability constitutes a significant portion of overall variability for per-section area measurements ($p = 0.026$), but the contribution of inter-observer variability is *not* significant for summed area measurements ($p = 0.146$). While the apparent proportion of total variability attributable to inter-patient effects has increased from the baseline measurements, it

ICC Statistic	Value	95% Confidence Interval
ICC_{pat}^{slice}	0.92	[0.888, 0.943]
ICC_{obs}^{slice}	0.002	[0, 0.109]
ICC_{pat}^{sum}	0.922	[0.865, 0.959]
ICC_{obs}^{sum}	0.002	[-0.001, 0.158]

Table 3.8: Intra-class correlation statistics for the follow-up area measurement data, calculated from Table 3.5.

should be noted that the patient population is the same and therefore the disease burdens in the different patients are just as variable as before. However, the reduction in total variability is largely due to the reduction in inter-observer variability, and therefore the relative contribution of inter-observer effects to total variability is considerably lower than at baseline.

The relative change between the baseline area measurements and follow-up area measurements was calculated, and the average bivariate rank correlation between observers for the change in area measurement from baseline to follow-up in summed area measurements was $\bar{\rho}^{\Delta sum} = 0.802$ ($p < 0.0001$). Using the WHO classification criteria (-50%/+25%), $\kappa = 0.557$. When classification criteria are used instead from geometries more specific to mesothelioma, such as the crescent (-79%/+120%), lens (-65%/+71%), and annulus (-81%/+163%) geometries from Oxnard *et al*. [51], $\kappa = 1.0$, $\kappa = 0.866$, and $\kappa = 1.0$, respectively. These values are summarized along in Table 3.9. The values of κ are fairly high for the Oxnard geometries due to the widening of the definition of stable disease (SD); for the annulus and crescent geometries, 30 of 31 patients were classified as stable disease, and only one patient was classified as progressive disease (PD), and there was universal agreement in the response categories that resulted from the three observers'

Response Classification Criteria	PR cut-point	PD cut-point	Generalized κ
World Health Organization	-50%	+25%	0.557
Oxnard *et al.*: Annulus	-81.2%	+163.0%	1.0
Oxnard *et al.*: Lens	-65.3%	+71.0%	0.866
Oxnard *et al.*: Crescent	-79.3%	+119.6%	1.0

Table 3.9: Generalized κ statistics for response classification performed using follow-up summed area measurements and a range of existing bi-dimensional response classification criteria.

contours. For the lens geometry response criteria, one patient was classified universally by all three observers' contours as PD, 28 patients were universally classified as SD, one patient was universally classified as partial response (PR), and for one patient, the contours from two observers resulted in a classification of PR while the contours from the third observer resulted in a classification of SD. Therefore, the vast majority of patients were classified as SD for the Oxnard geometries, and there was good agreement among observers for the few patients classified as either PR or PD.

3.4 Discussion

The purpose of this study was to assess manual area measurements of malignant pleural mesothelioma as a metric for response evaluation. Specifically, in the baseline study the area measurement technique itself was investigated to determine whether such measurements would be well-suited for response evaluation, and it was discovered that substantial variability was attributable to inter-observer effects. The measurement technique was evaluated both in terms of individual section-by-section measurements and in terms of the sum of three sectional area measurements per patient.

These summed measurements are more clinically relevant, just as summed linear measurements across three CT sections are currently used in the modified RECIST system.

In the follow-up portion of this study, the variability of area measurements was assessed in a more practical context, since in clinical practice response measurement acquisition involves reference to previous measurements. The variability of area measurements (both per-section and summed measurements) may be large when each observer is given a "blank slate," as in the baseline study, but clinical practice involves only a single individual making baseline measurements. Follow-up measurements are made based on this existing baseline measurement, and therefore the baseline measurement serves as a biased rubric for a follow-up observer (implicitly constraining the thought processes and actions of the observer, leading to a reduction in inter-observer variability).

The results for baseline area measurements indicate that the 95% confidence intervals of relative inter-observer variability are considerably larger than the reasonable two-dimensional response classification criteria found in Table 3.9. Since the use of a 95% (i.e., 2σ) confidence interval may seem overly conservative, the calculations for relative inter-observer variability can be altered to yield the 68% (i.e., 1σ) confidence interval by changing the quantity 1.96 to unity in equation 3.5. With this change, the 68% confidence interval for relative inter-observer variability of section-by-section area measurements is [-46%, +84%], and for summed area measurements the same interval is [-46%, +87%]. Even these 68% confidence intervals are so large as to approach the classification criteria in Table 3.9.

The results of Table 3.3 and Figures 3.3 and 3.4 support the concept that the observers had different ideas of what constitutes identifiable disease. If the observers had simply been sloppy in their measurements, one would expect a reduction in inter-observer variability from the per-section measurements to the summed measurements. Random fluctuations on one slice for a given patient might be canceled from variability on the second and/or third slices, leading to higher variability estimates for per-section measurements. The results of the baseline study, however,

indicate that the inter-observer variability is nearly identical between per-section and summed area measurements, leading to the conclusion that the observed variation results from different contouring styles (for lack of a better term) between observers, which are preserved from the per-section measurement setting through to the summed measurements. One possible explanation for the different contouring styles between observers is that area measurements are not part of the standard clinical workflow and therefore the observers are not as familiar with the technique as they are with, say, the linear measurement technique.

As expected, the follow-up estimates for inter-observer variability are reduced from the baseline estimates (compare Table 3.7 with Table 3.3). The fact that the variability was reduced is not noteworthy; previous studies have shown that exposure to prior contours heavily biases observers [92], and therefore only the absence of any reduction in variability from baseline to follow-up would have been noteworthy. However, the variability is reduced to the point where inter-observer variability is no longer a significant component (only 0.2%) of total variability in summed follow-up area measurements. The relative inter-observer variability is still fairly wide, with a 95% confidence interval of [-41%, +69%] (even the 68% confidence interval for relative inter-observer variability of summed follow-up area measurements is [-24%, +31%]). Even these implicitly constrained follow-up measurement variabilities are on the same level as the reasonable classification criteria given in Table 3.9, and therefore tumor response could be misclassified due to observer variability effects alone.

While the inter-observer variability in follow-up measurements is clearly reduced from baseline measurements, the remaining amount of variability is not zero. Part of this variability may be due to differences in disease presentation between baseline and follow-up; a small number of scans had a difference in contrast administration between baseline and follow-up, and other scans may exhibit differences in patient orientation in the CT scanner bore. Any misalignment in patient yaw would result in difficulties for matched-slice selection at follow-up, and one observer noted this challenge on a small number of patients. In regular clinical practice, when the disease

presentation changes for non-physiological reasons (e.g., what may be obvious vasculature with a contrast-enhanced follow-up CT scan may have previously appeared to be pleural disease if the baseline CT scan had been acquired without contrast), the observer can note the change in the radiology report, and future response assessment measurements can be made according to the changed interpretation. In this study, however, observers were asked to make follow-up measurements of the disease that had previously been outlined on the baseline scan. For the patient cases in which disease presentation was changed from baseline to follow-up in part due to image acquisition, different observers may have had different interpretations and contouring strategies. This ambiguity in how to handle changes in disease presentation is a source of variability in clinical practice, and its implicit inclusion in this analysis is important for the applicability of this study's conclusions.

It is important to note that the sets of two-dimensional response classification criteria used in the calculation of the κ statistics are somewhat arbitrary in that no one set of the criteria are known to be appropriate for mesothelioma. The original WHO criteria are given for bi-dimensional measurements, which are not the same as full area measurements, and the same is true for the theoretical geometric models derived in the Oxnard *et al.* paper. The interesting difference between the WHO and the Oxnard criteria is the substantial widening of the definition of the stable disease category. When more tumor responses are swept into this SD categorization, there will be a corresponding increase in any concordance statistic (such as κ), and this is observed in the results above.

Of course, our study is limited by the number of patients and radiologists available for analysis. The relative inter-observer variabilities presented herein should not be taken as evidence that area measurements for MPM are a "dead end." Studies with more patients and/or radiologists could potentially provide updated estimates of inter-observer variability. However, the collection of these data is quite time consuming.

Even if follow-up variability were reduced further from the results given above, it would be hard to recommend the adoption of area contours as a tool for response assessment. The observers in this study reported time durations for contouring the disease area on three axial slices on the

order of 20 minutes, and as we will see in subsequent chapters, this is approximately the time required to complete the manual intervention necessary for a full volumetric semi-automated disease segmentation. The volumetric segmentations will capture complete tumor burden changes and will likely exhibit very low inter-observer variabilities. For instance, in the study by Frauenfelder *et al*. [57], a value of $\kappa = 0.9$ was reported for volume measurements classified according to a spherical geometric model, while in this study, $\kappa = 0.557$ for area measurements classified by a spherical geometry model. Given the inter-observer variability estimated for area measurements in this study and the time required to make such area measurements, it is difficult to foresee a situation in which summed area measurements (pseudo-volumes) would be used in place of semi-automated segmentations of actual disease volume for response assessment.

CHAPTER 4

DISEASE VOLUMES AS A MARKER FOR PATIENT RESPONSE

"The wheel is come full circle." – William Shakespeare, *King Lear*

4.1 Introduction

Any image-based response evaluation method has two components; the first describes a protocol for making measurements, and the second describes how to classify patients into response categories once those measurements are available. Since 1981, tumor response assessment with medical images has been focused on reducing the dimensionality of the first component and discretization of the second component. The original World Health Organization (WHO) bi-dimensional measurement technique focused on the product of two linear measurements, the longest diameter and longest perpendicular measurement to said diameter [42]. The product of these measurements was used as a quasi-two-dimensional metric to assess tumor response by classifying progressive disease (PD) for an increase of more than 25% from the minimum of previous measurements (nadir) and partial response (PR) if the measurement decreased by 50% or more from the baseline measurement. Measurements not meeting the criteria for either PD or PR were classified as stable disease (SD). These -50%/+25% cutoffs for PR and PD, respectively, were based on previous breast cancer cohort studies, where the origin of the criteria was, in part, from what physicians believed to be reasonable palpable changes in tumor burden (i.e., the amount of change the physicians could reliably identify from palpation alone).

Later, the Response Evaluation Criteria In Solid Tumors, or RECIST criteria, were derived to simplify this measurement process; the two necessary measurements were reduced to one (again, the longest tumor diameter), and the cutoff criteria were derived from a geometrical relationship [43]. For a sphere where the cross-sectional area increases by 25%, the diameter increases by 12%, whereas a cross-sectional area reduction of 50% for the same sphere would correspond

to a diameter decrease of 29%. The -29%/+12% "sphere model" extrapolations were rounded to $-30\%/+20\%$, leading to the current RECIST classification criteria [44, 45]. The RECIST measurement technique and classification criteria are currently used across many disease types including lung cancer, mesothelioma, breast cancer, colorectal cancer, prostate cancer, gastro-intestinal stromal tumors, soft tissue sarcomas, brain tumors, renal cell carcinomas, and others [46].

With the advent of near-isotropic three-dimensional voxels in medical imaging and advances in computer vision methodology, there has been a more recent drive to use full three-dimensional volume measurements for response assessment [45, 104, 105] and to track measurements over time using continuous, not discretized, response [44, 106]. Especially for diseases such as mesothelioma, where the morphological characteristics do not correspond to the spherical extrapolation of the RECIST classification criteria, disease volumes are a logical choice for tumor burden assessment. While the past history of response assessment has moved away from volumetric quantification for reasons such as imaging technology and ease of manual measurement, the previous techniques used for response assessment were always intended to identify tumor burden changes (i.e., volumetric changes), and now is the time for the wheel to come full circle (to paraphrase Shakespeare).

Numerous studies have investigated the use of tumor volume and tumor volume-related measurements for response assessment in patients with MPM using MRI, CT, and FDG-PET imaging [35, 37, 54–58]. The main challenge in these volumetric studies is the segmentation of the complete tumor volume. Using FDG-PET imaging, the segmentation of MPM is greatly facilitated by the FDG avidity of the tumor. Three-dimensional semi-automated region growing techniques can be used with high reproducibility to capture the complete tumor volume [55, 59], and even more rudimentary simple thresholding techniques also appear to perform well [56].

The segmentation of MPM in CT images is more problematic, and both published studies use a semi-automated tool for MPM volume segmentation [57, 58]. The study by Frauenfelder *et al.* [57] used a linear shape-based interpolation technique, requiring contours on "every fourth

or fifth slice." The main conclusion of this study related to volumetric response was that the inter-observer agreement of volumetric response classification is much higher than for manual modified RECIST response classification (general $\kappa = 0.9$ vs general $\kappa = 0.33$, respectively) [57]. The study by Liu *et al.* [58] utilized a combination of semi-automated techniques for volumetric MPM segmentation, and their analysis revealed changes in tumor volume to be significantly associated with patient survival. Patients experiencing tumor growth had a median survival of 11.5 months, while patients with tumor decrease had a median survival of 18.1 months.

Of course, tumor burden is not the only factor affecting a given patient's outcome during or following therapy. Other covariates may provide crucial information useful for the prediction of survival and have been shown to do so in patients with MPM. Section 1.1.4 outlines covariates used previously in patients with MPM, including patient age, sex, performance status (PS), white blood cell (WBC) count, blood platelet count, histological cellular subtype, presence of chest pain, and weight loss [30–35]. In fact, some of these covariates (and others) have been shown to be significant independent predictors of patient survival in models that also include disease volumes estimated from FDG-PET imaging [35].

The goal of this study was to create a comprehensive model to predict MPM patient survival utilizing time-changing estimates of disease volume in conjunction with other clinical covariates. To date, no survival model for MPM utilizes disease volumes that can freely change over time. Continuous measurements of disease volume were hypothesized to be a significant predictor of MPM patient survival, both as a single covariate and in conjunction with other clinical covariates.

4.2 Patients and Methods

4.2.1 Patient Cohort

The patient cohort used in this study was a subset of the entire "Perth database" described in section 2.2.1 and Table 2.1. For inclusion in this specific study, patients were required to have

an available CT study at baseline (prior to beginning chemotherapy) and one or more follow-up CT scans during (and immediately following) chemotherapy. This constraint reduced the eligible patient cohort to 81 patients, and the summary description of these patients is given in Table 4.1.

Table 4.1: Description of the patient cohort used in this specific study, consisting of 81 of the original 97 patients. This specific patient cohort is a subset of the patients summarized in Table 2.1.

Characteristic	Summary
Sex:	
Male	n = 68
Female	n = 13
Age at Diagnosis:	
Median	66 years
Range	41–80 years
Chemotherapy:	
Carboplatin/Pemetrexed	n = 7
Cisplatin/Pemetrexed	n = 42
Cisplatin/Gemcitabine	n = 32
Histology:	
Epithelioid	n = 60
Sarcomatoid	n = 7
Biphasic	n = 14
T Stage:	
T1	n = 19
T2	n = 22
T3	n = 25
T4	n = 15

(continued on next page)

(Table 4.1, continued from previous page)

N Stage:		
	N0	n = 25
	N1	n = 3
	N2	n = 40
	N3	n = 13
M Stage:		
	M0	n = 70
	M1	n = 11
IMIG Stage:		
	I	n = 13
	II	n = 5
	III	n = 36
	IV	n = 27
Known Asbestos Exposure:		
	Yes	n = 75
	No	n = 6
Chest Pain:		
	Yes	n = 50
	No	n = 31
Shortness of Breath:		
	Yes	n = 67
	No	n = 14

(continued on next page)

(Table 4.1, continued from previous page)

ECOG Performance Status:		
	0	n = 38
	1	n = 38
	2	n = 5
Talc Pleurodesis:		
	Yes	n = 30
	No	n = 51
Weight:		
	Median	75 kg
	Range	52–121 kg
Height:		
	Median	172 cm
	Range	155–189 cm
Smoking Status:		
	Never	n = 36
	Past	n = 39
	Present	n = 6
Pleurectomy/Decortication:		
	Yes	n = 1
	No	n = 80

4.2.2 Imaging

Patients were imaged using helical CT up to one month prior to the first cycle of chemotherapy and throughout their treatment regimen (typically after the first cycle, then every two cycles thereafter).

CT staging was performed according to the Union for International Cancer Control (UICC) TNM staging system (2002). CT scans were staged by a thoracic radiologist or medical oncologist experienced in mesothelioma imaging.

There were a total of 281 CT scans in this study, with a median of four scans per patient (including baseline scans). Ten patients had only a baseline scan with one follow-up scan, while 30 patients had three scans total, 34 patients had four scans total, and seven patients had five scans total. The median duration between scans was 49 days. Of the 281 scans, 197 were performed on General Electric scanners (HiSpeed CT/i, n=105; LightSpeed Pro 16, n=2; or LightSpeed VCT, n=90), and 84 were performed on Philips Brilliance 64-slice scanners. At least 135 of the scans were performed with iodinated contrast media (for the other 146 scans, the contrast field in the DICOM image header is empty, which does not necessarily imply that no contrast was administered).

Only one reconstructed series was required for each CT scan date, and this series was selected for each patient primarily with consideration for slice thickness and secondarily with consideration for reconstruction kernel. Preference was given to thinner slice thicknesses and "Standard" reconstruction kernels, but if for a given patient there was a scan date with only "Lung" kernel reconstructions, then matched kernel and slice thickness reconstructions were used for the other scan dates. Having this type of consistency across the scan dates for a given patient was considered important for segmentation of volumetric disease, since different amounts of disease might be segmented on different reconstructions due to, for instance, partial volume effects. With this consistency in mind, slice thickness for the different scans in this study was 0.63 mm (n=6), 1 mm (n=18), 1.25 mm (n=36), 2.5 mm (n=97), or 5 mm (n=124). In-plane voxel dimensions ranged from 0.54–0.87 mm, and all reconstructed axial images had an in-plane matrix size of 512 by 512. The kVp setting for the scans was predominantly 120 kVp (n=273), with 100 kVp (n=2) and 140 kVp (n=6) also used. Reconstruction kernels fell into two broad categories, with "Lung" kernels (including the Philips "L" and GE "Lung" kernels) used for 185 scans and "Standard" kernels

(including Philips "B" and GE "chest," "soft," and "standard" kernels) used for the remaining 96 scans.

4.2.3 Disease Segmentation

For each scan, the pleural disease was segmented using a semi-automated method based, in large part, on the Ph.D. thesis of William Sensakovic at The University of Chicago [60, 95]. Because of the considerable overlap in Hounsfield Unit (HU) values between actual mesothelioma tumor and pleural effusion, the automated method used by Sensakovic produces contours of pleural disease and does not readily separate tumor from effusion. This is discussed further below. Therefore, the end goal of the disease segmentation technique used in this study was reliable volumes of pleural disease and not necessarily volumes of *only* mesothelioma tumor. Note that since there is no database of "gold standard" pleural disease volume segmentations (it would be prohibitively time-consuming to generate such segmentations), the existing automated segmentation method was only partially validated by Sensakovic *et al*. [60]; the automated segmentation technique was found to agree with human observers on a limited number of individual axial CT slices only to the extent that human observers agreed with one another.

4.2.3.1 Automated Segmentation Methods

The first step in the segmentation method is a thoracic segmentation. The thoracic segmentation identifies just the patient in the image, removing the CT scanner table and any blankets on top of the patient. Identifying the patient in the image is an important factor in guiding the rest of the segmentation methods. Following the thoracic segmentation, a bone and contrast segmentation identifies regions of high HU values using a thresholding technique followed by morphological operations to remove small regions and fill the marrow "holes" in the rib cross-sections. An airway segmentation is calculated next, including the trachea and bronchi. This segmentation is performed three-dimensionally using automatically seeded active contours and is necessary because of the

similarity in HU values between lung and airway regions. Without this segmentation, the lung segmentation could potentially "leak" into the main-stem bronchi and trachea.

Lung region segmentation is a crucial component of the disease segmentation process. Generally speaking, any "disease" is located between the visceral and parietal pleural surfaces, where the parietal pleura is the outer surface in contact with the chest wall and the visceral pleura is the inner surface attached to the lungs and other hemithoracic structures (e.g., vasculature and mediastinum). Therefore, the lung segmentation helps define the "inner" boundary of any potential pleural disease. The lung segmentation method is fully automated and utilizes gray-level, morphological, and texture features to segment the aerated lung tissue. The lung segmentation method has been under development in our group for some time, and the updated version been used for other studies for patients with MPM [107, 108]. Figure 4.1 shows the above-mentioned automated segmentation results for an example patient (the same patient scan is used to exemplify all the segmentation methods described herein).

Following lung segmentation, a hemithoracic segmentation is necessary to define the "outer" pleural surface. The method used in reference [60] is fully automated and relies on an active contour region growing approach to propagate the lung contours outward until they are constrained by the hemithoracic bounding structures, such as the ribs and mediastinum (as partially defined by the bone and contrast segmentation mentioned above). Unfortunately, this technique is prone to errors, such as the active contour being "pulled through" the openings between the ribs, especially near the scapulae and clavicles. After initial review of the automatic hemithoracic segmentations for a large number of applicable patient scans, it became clear that the previously published method would not yield the necessary segmentations on the new (and potentially more diverse) patient scan cohort (see Figure 4.2 for an original automated hemithoracic segmentation). The cohort of patient scans used to build the original algorithm had a tendency toward thinner slice thickness than the patient cohort used in this study, and the level of anatomic variability in MPM patients (rib bunching, surgical meshes, etc) complicates any attempted automated method.

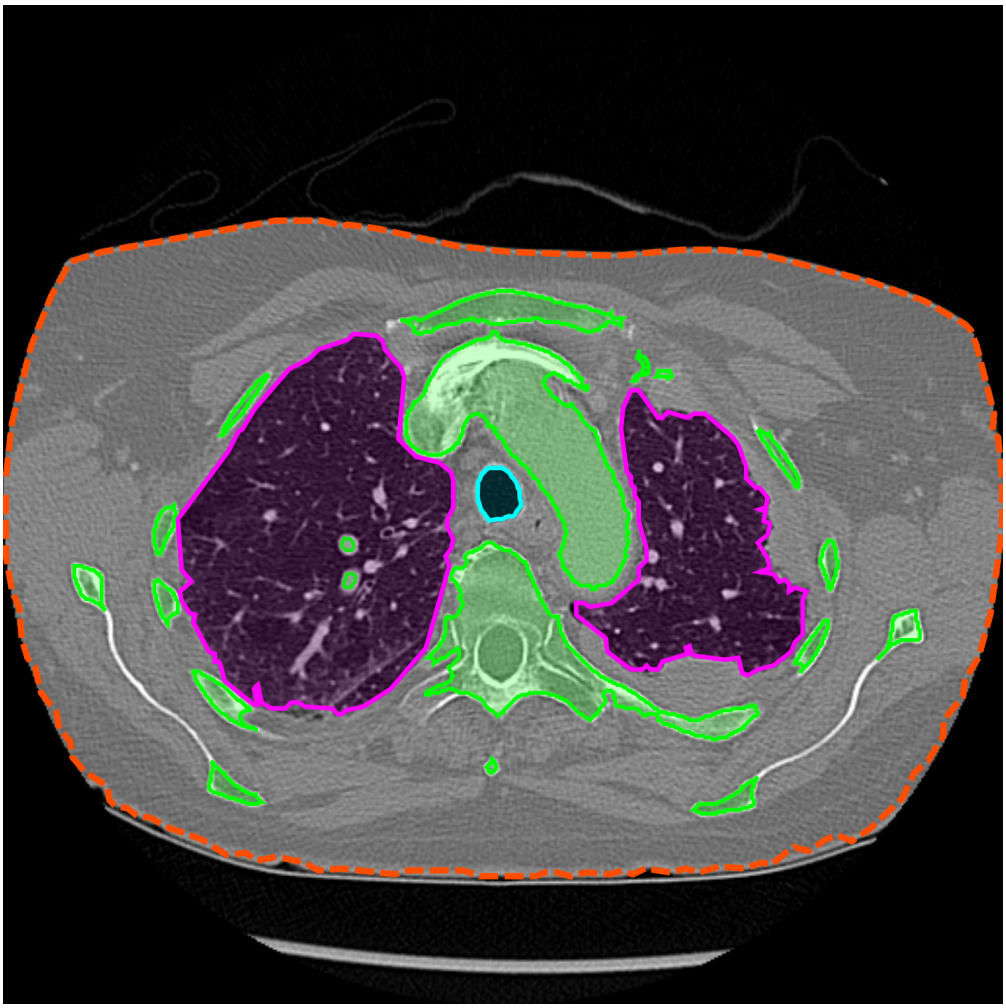


Figure 4.1: Automatically segmented components of an example patient scan, including the thoracic segmentation (dashed orange), bone and contrast segmentation (green), airway segmentation (blue), and lung region segmentation (pink).

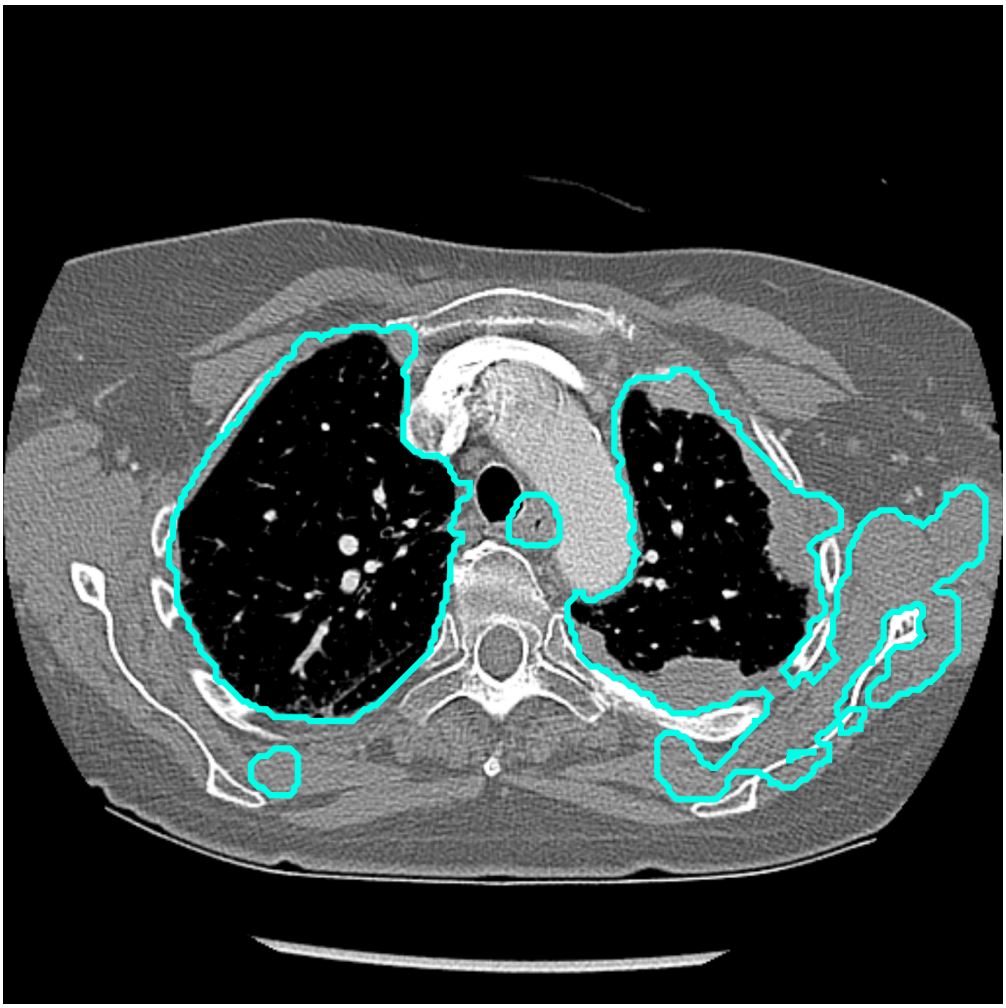


Figure 4.2: Results of automated hemithoracic segmentation from reference [60] on example patient shown in Figure 4.1. Note how the contour in the left hemithorax has been "pulled through" the opening in the ribs to encompass a large portion of musculature.

In an attempt to improve the automated hemithoracic segmentation algorithm, efforts were made to develop an automated rib segmentation algorithm. With a rib segmentation algorithm, one could constrain the hemithoracic segmentation to be "inside" the ribcage (some patients exhibit invasion of disease through the parietal pleura into the chest wall, and in such instances the invasive disease foci would be erroneously excluded). Automated rib segmentation on CT images is a task that has been attempted by other groups, with apparent success [109–112]. The methodology in two of the previous algorithms was implemented to recreate a functional algorithm (the other two were developed by corporate researchers, and methodology was not readily available).

The method starts with an automated spinal canal segmentation inspired by Lee *et al.* [111] derived from a distance mapping from the bone segmentation mentioned above. Next, the bone segmentation was enhanced using a Hessian filter implemented in the Insight ToolKit (ITK), as inspired by Staal *et al.* [110]. For each voxel in the volumetric image block, the Hessian tensor H is calculated as

$$H = \begin{pmatrix} L_{xx} & L_{xy} & L_{xz} \\ L_{yx} & L_{yy} & L_{yz} \\ L_{zx} & L_{zy} & L_{zz} \end{pmatrix},$$

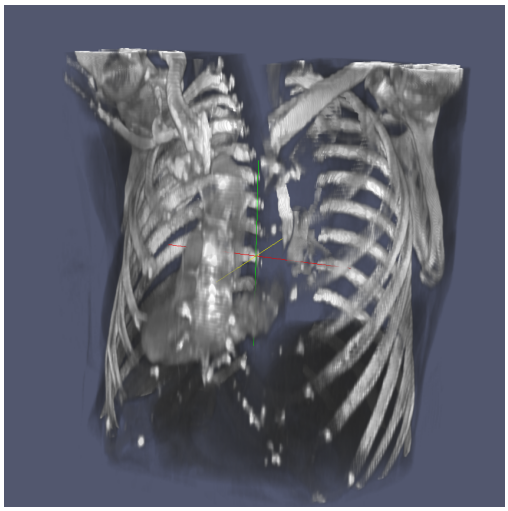
where L_{ij} is the second-derivative of image intensity values as calculated using a Laplacian-of-Gaussian (LoG) filter. For "tubes," the Hessian tensor will have two eigenvalues of relatively large magnitude, and one eigenvalue of small magnitude. The eigenvector associated with this last eigenvalue will point along the direction of tube curvature in the original image. The ITK implementation of the Hessian filter allows for automatic enhancement of "tubularity" after performing enhancement at multiple spatial length scales [113]. Following Hessian enhancement, rib "seeds" were identified using the proximity to the spinal canal segmentation, as in [111]. The image block was processed using a sequence of morphological operations (erosion and opening) in MATLAB (MathWorks; Natick, MA) in order to separate rib regions from areas of contrast in the mediastinum [114]. The $\vec{z}$-component of the principal curvature eigenvector from the Hessian tensor

computation was used to remove enhanced regions that were primarily oriented in the superior-inferior direction, such as the edges of the scapulae and patient vasculature. The resulting regions were wrapped in a 3D convex hull in an attempt to identify the outer extent of the ribcage. The process is illustrated in Figure 4.3.

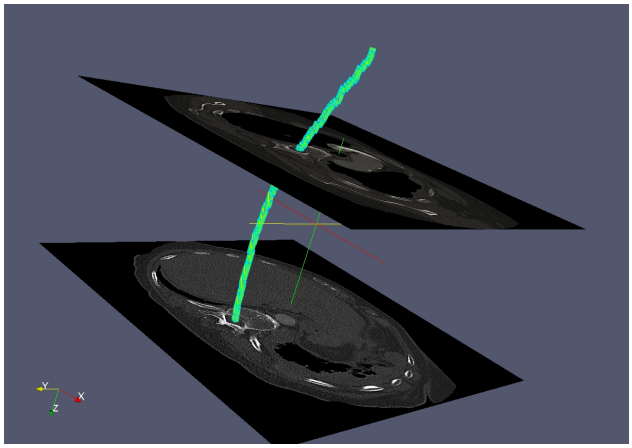
After implementation of this simplistic rib segmentation algorithm, the results from the patient cohort in this study indicated that the method would not be viable without a substantial amount of further work on the topic. Many issues arose that complicated the automatic extraction of *just* the ribcage. For instance, differences in patient contrast administration greatly affected the relative brightness of different structures in the body. Figure 4.4a shows a patient scan where a substantial amount of contrast agent remains in the subclavian vein. Since this vein comes in close contact with the clavicles and first rib, it was impossible to separate the vascular structure from the bone structures without removal of the bone structures themselves. Therefore, the wrapped hull was artificially pulled anterio-laterally. Other problems arose due to rib bunching and implanted surgical meshes, which alter HU values in the CT scan and thereby affect the Hessian enhancement and rib segmentation. Undoubtedly, the ribcage segmentation algorithm could have been improved substantially with more time and a training database with defined "truth," for instance through the inclusion of a candidate-region classifier [110]. Even so, the rib segmentation technique somewhat improved the previous automated hemithorax segmentation, where the updated hemithorax segmentation is shown in Figure 4.4b for the same patient as in Figure 4.2.

While the inclusion of the rib hull segmentation in the automated hemithoracic segmentation algorithm appears to improve the hemithorax contours (compare Figure 4.4b to Figure 4.2), the resulting pleural disease segmentations were still too far from "usable" to use in this study. The pleural disease segmentation approach was exactly as described in [60], but a brief summary is in order. The pleural space was defined as everything simultaneously "inside" the hemithoracic segmentation and "outside" the lung segmentation, or more formally,

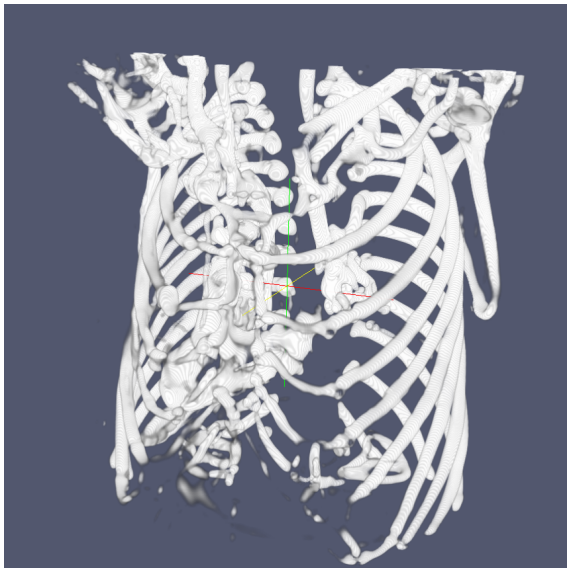
$$\text{Pleural Space} = (\text{Hemithorax}) \bigcap (\text{Lung})^{c}, \tag{4.1}$$



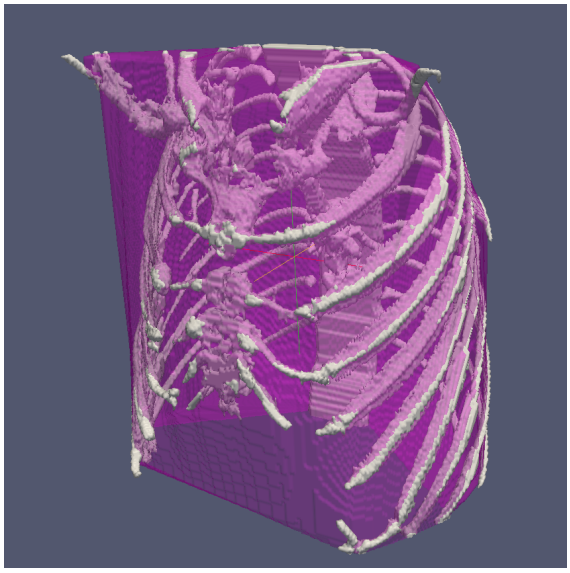
(a)



(b)

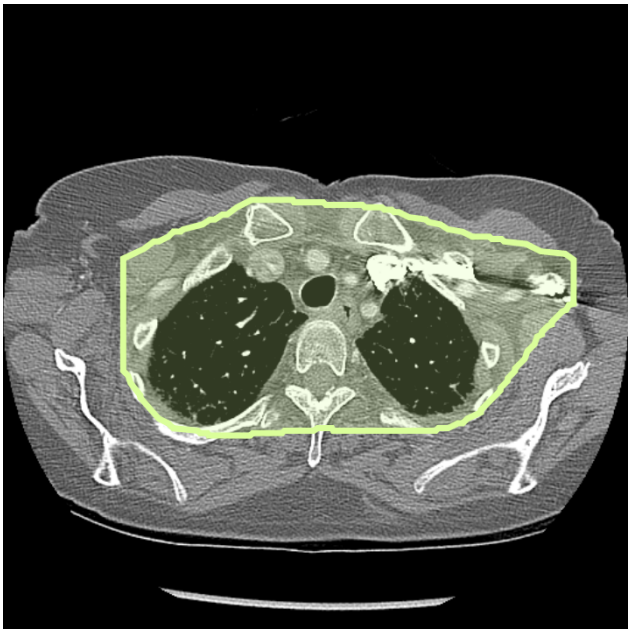


(c)

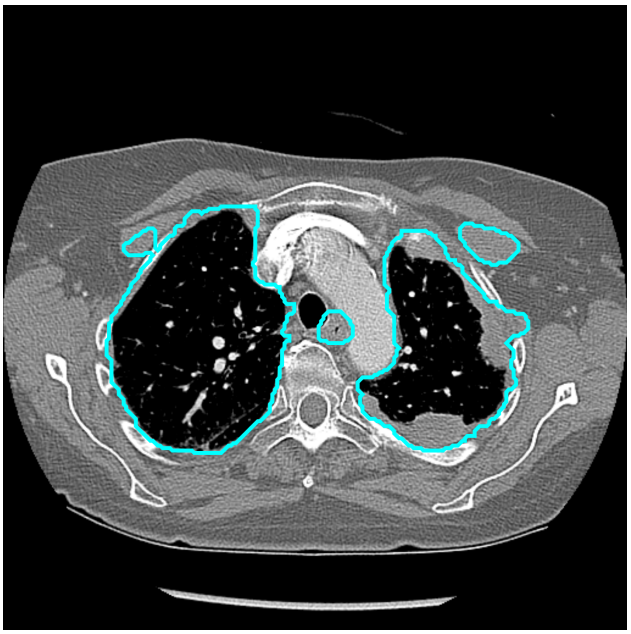


(d)

Figure 4.3: Attempted automated rib segmentation algorithm, constructed from published methodology [110, 111]. 4.3a, volume rendering of chest CT scan. 4.3b, output of automatic spinal canal segmentation. 4.3c, hessian-enhanced chest CT. 4.3d, final output, with the ribcage segmentation in white and the 3D "wrapped" hull used as the hemithorax boundary in pink. Note the persistence of carotid and subclavian arteries, which erroneously extend the hull anterio-laterally.



(a)



(b)

Figure 4.4: 4.4a, rib hull segmentation, showing the effect of remnant contrast in the subclavian vein. 4.4b, results from the automated hemithoracic segmentation algorithm after the inclusion of the rib hull segmentation to define an outer bounding structure. Both images are of the same patient scan as in Figure 4.2, though 4.4a is not the same axial section as shown previously.

where $()^c$ indicates the complement of a set. A k-means classifier is applied to the diffusion-filtered image values in this initial pleural space segmentation, where the classifier is initiated with nine classes centered at HU values of -10000, -500, -200, -50, 0, 50, 100, 200, and 500 (for extrapleural space, aerated lung, air/lung/fat, fat/soft tissue, soft tissue/mesothelioma/effusion, soft tissue/mesothelioma/effusion, soft tissue/mesothelioma/cartilage, mesothelioma/cartilage/bone, and bone/metal/contrast media, respectively). Pixels assigned to the first two or last categories after iterative convergence of the k-means classifier are eliminated from the segmentation. Finally, a watershed segmentation is applied to the resulting segmentation mask [115]. The watershed segmentation splits the pleural space segmentation into three-dimensional regions based on their morphology and proximity to one another, and this step is performed only to expedite the process of final manual segmentation editing (an entire three-dimensional labeled region can be removed with the click of a button). The pleural space segmentation is shown in Figure 4.5. The watershed segmentation would allow for quick removal of many of the erroneous regions in the image (such as the regions indicated by red and yellow arrows), but other errors are combined with true disease (such as the green and magenta arrows). Editing the three-dimensional segmentation shown in part by Figure 4.5 required approximately 45 minutes by a trained user.

4.2.3.2 Semi-Automated Techniques

In order to obtain pleural disease segmentations in a more reasonable amount of time, a semi-automated hemithoracic segmentation algorithm was developed. The semi-automated approach utilized shape-based interpolation [116] and an interface written for the in-house software Abras. The user places manual contours of the hemithorax spaced axially approximately every 15 mm, depending on the local level of hemithorax boundary change. After contour placement on the selected slices, the uncontoured axial sections are "filled in" automatically with contours interpolated from the initial contours. The user is then allowed to change any contour necessary, iterating the interpolation process (any time a contour is changed, it is automatically defined as a new seed

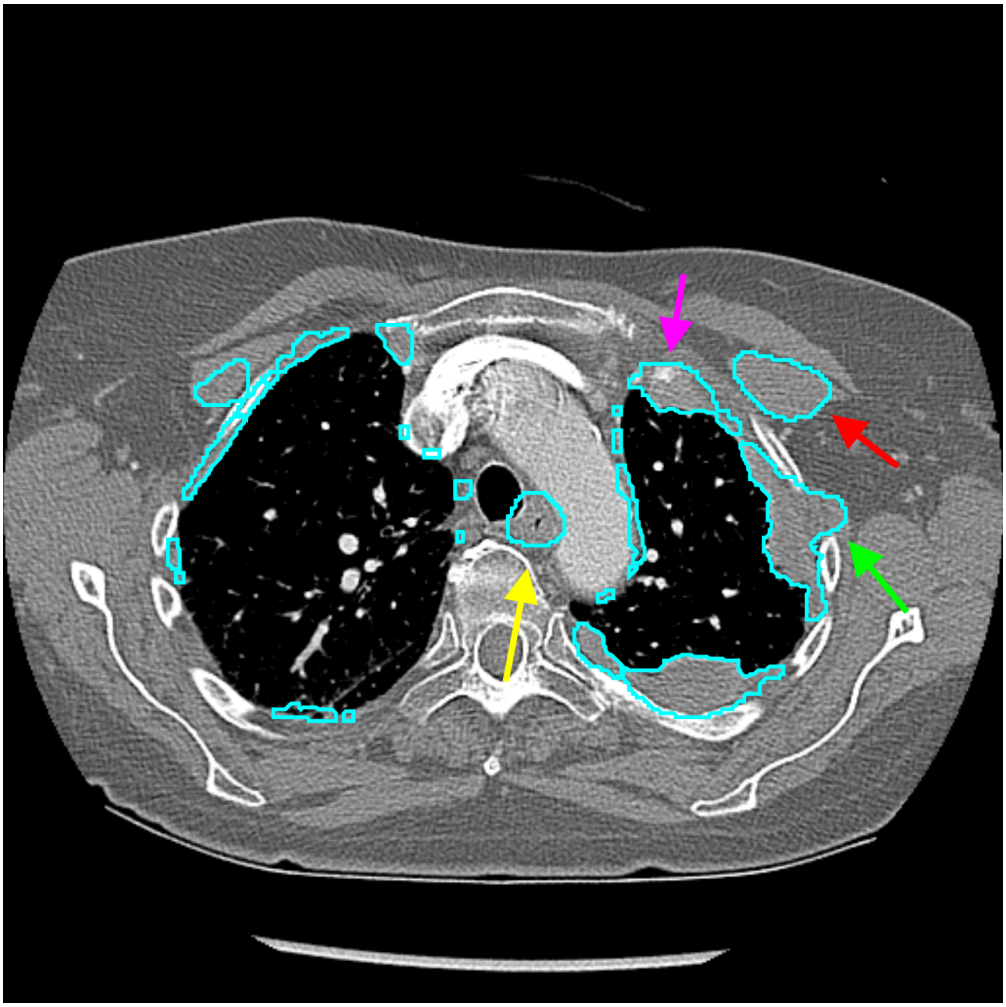


Figure 4.5: Automated pleural space segmentation from hemithoracic segmentation in Figure 4.4b. Erroneous regions indicated by arrows. The magenta arrow indicates inclusion of the end of a rib. The green arrow indicates inclusion of neighboring chest wall musculature. The yellow arrow indicates the inclusion of the esophagus. The red arrow indicates chest musculature. Only the patient's left hemithorax contains disease.

contour for the next interpolation step). Once the resulting hemithorax segmentation is acceptable, the contours are saved.

The important distinction with this semi-automated method is that the user only needs to contour the outside boundary of any identifiable pleural disease. If the disease exists in only one hemithorax, the user only needs to make contours on one side of the patient. If the disease does not span the complete axial extent of the chest, the user only needs to supply contours in the applicable region. Whenever the user is able to reduce the contouring burden by focusing seeding efforts on a limited spatial region of interest, the time required to complete the semi-automated hemithorax segmentation will be reduced compared with the time required for seeding the full axial span of both hemithoracic spaces. An additional advantage of the semi-automated technique is that any instances of disease invasion through the chest wall can be manually included during the contour seeding process. The semi-automated hemithorax segmentation is shown in Figure 4.6.

After obtaining the semi-automated hemithorax segmentation, the pleural space segmentation proceeds in the automated fashion outlined above. The watershed segmentation is again applied to the resulting volumetric segmentation, and the in-house software Abras is used to manually edit and finalize the segmentation. Hemithorax segmentation seeding and pleural disease segmentation editing were performed by the author, who had been trained in thoracic anatomy by an attending radiologist (Christopher Straus, MD). Plans were initially made for the attending radiologist to review all of the 281 volumetric pleural disease segmentations, but after review of approximately 20 scans resulted in no changes to the contours of the author, these plans were abandoned as too time-intensive. The initial pleural disease segmentation and the final edited results are shown in Figures 4.7a and 4.7b, respectively.

To calculate pleural disease volume for each patient scan, a pixel-counting technique was used [94]. The number of pixels inside the pleural disease segmentation were counted for each axial section and multiplied by the area of a single pixel and by the axial section thickness. These per-section disease volumes were summed for each axial section containing disease; very occasionally,

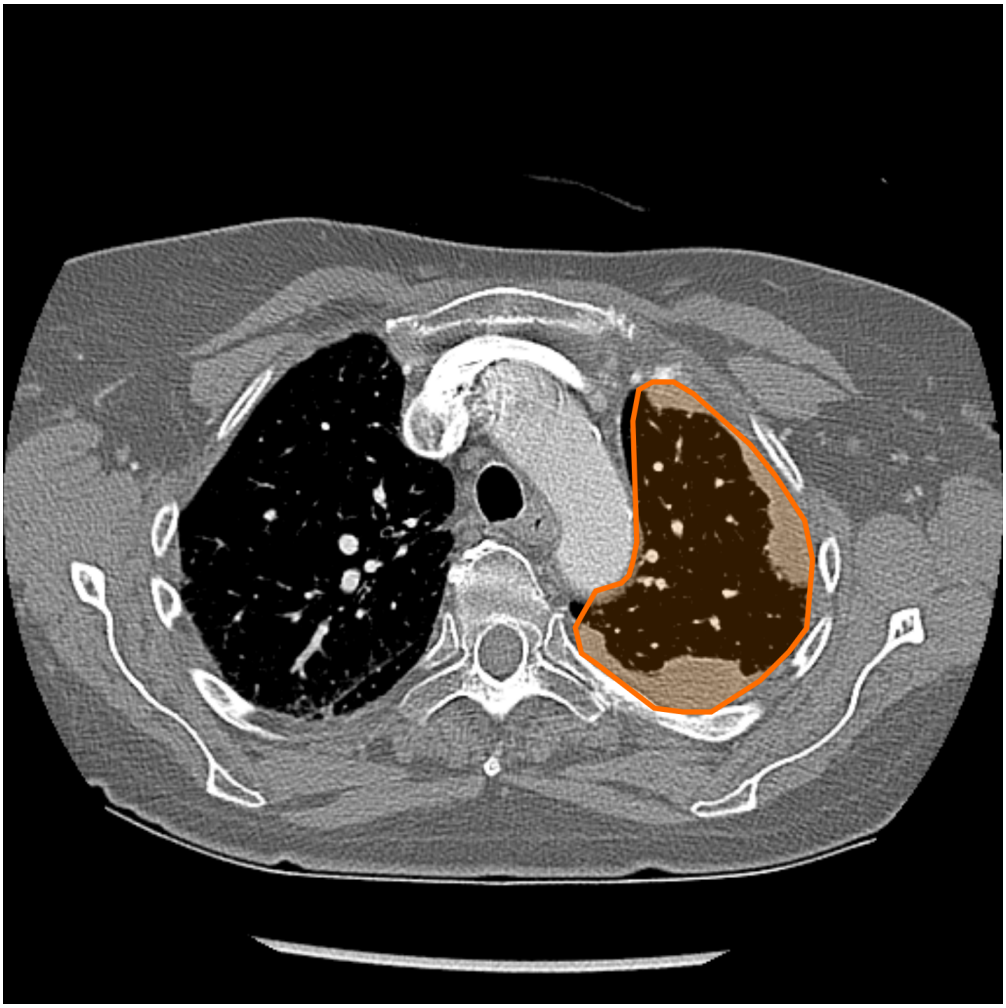
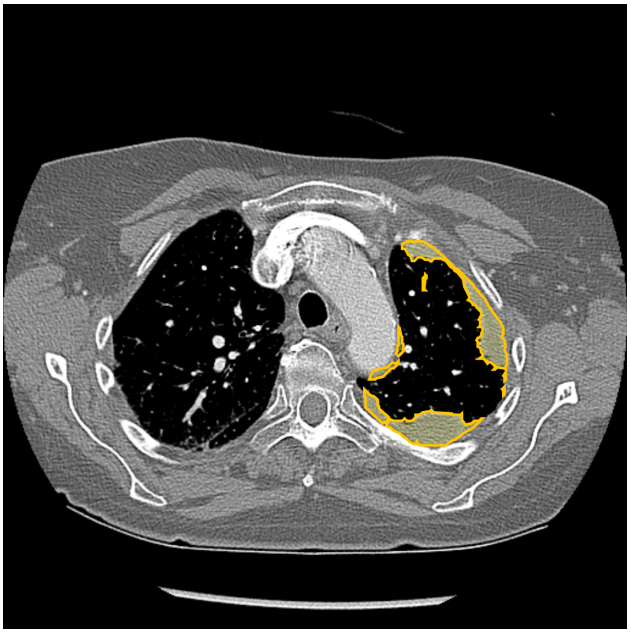
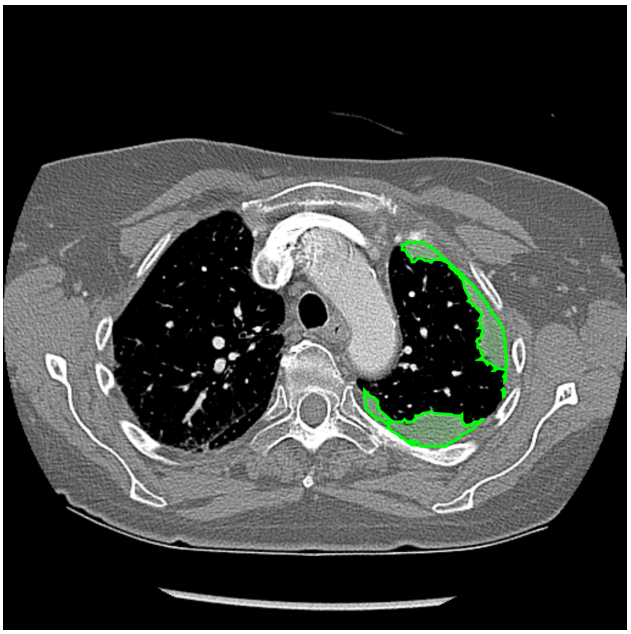


Figure 4.6: Semi-automated hemithorax segmentation from same patient section shown in Figure 4.2. This axial section did not contain a seeded contour, so the orange contour is an interpolation from other manual axial contours. Note that the hemithorax contour only needs to encompass the identifiable pleural disease.



(a)



(b)

Figure 4.7: Pleural disease segmentation from the semi-automated hemithorax segmentation (4.7a), and final edited pleural disease segmentation (4.7b). Because of the constrained semi-automated hemithorax segmentation, manual editing of pleural disease segmentations was generally minimal (restricted to partial volume artifacts adjacent to aortic arch in this axial section). The axial section is the same shown in Figure 4.6.

the axial section thickness changed over the span of a single scan, and therefore the volume of disease needed to be summed on a section-by-section basis. Further complicating the issue was the sometimes incomplete axial coverage of a single scan (for instance, when the disease extended inferiorly past the most inferior axial section). To account for this constrained scanning, the disease volume was calculated for each patient only in the anatomical portion of the thorax shared across all patient scan dates. The matching of the inferior limit axial section was performed manually by referencing vertebral bodies and rib origin points whenever possible, but occasionally with additional reference to organ location when respiratory phase was prohibitively disparate between scan dates. This manual matching of inferior limits was required in approximately one-quarter of the scans used in this study.

To quantify overlap between two segmentation techniques of the same structure (producing contours denoted by, say, Seg_1 and Seg_2), the Jaccard similarity coefficient J provides a useful metric. Also known as the area of overlap metric (AOM), J is defined as

$$J = \frac{\text{Volume}\,(Seg_1 \bigcap Seg_2)}{\text{Volume}\,(Seg_1 \bigcup Seg_2)}, \tag{4.2}$$

where the numerator indicates the volume of the intersection of the two segmentation techniques, and the denominator indicates the volume of the union of the two segmentation techniques. For segmentations with no overlap whatsoever, $J = 0$, and $J = 1$ for segmentations that overlap completely. In this study, J was only truly applicable to pleural disease segmentations, since although multiple methods were used to identify the hemithoracic boundary, their aims were so different as to make direct comparisons misleading (e.g., the automated techniques produce bilateral segmentations, while the semi-automated technique produces segmentations bounding the identifiable disease). The value of J was calculated to compare the automated pleural disease segmentations with the semi-automated pleural disease segmentations, and also to quantify the extent of manual intervention in the pleural disease segmentation editing process, comparing the pre-edited pleural disease segmentations with the final edited pleural disease segmentations.

4.2.4 Survival Analysis

As mentioned in Chapter 1, the most common method used to model patient survival and assess the correlation between covariates and survival is Cox proportional hazards (PH) modeling [61, 66]. The standard PH survival model assumes that the effect of any given covariate is fixed across time and the hazard $h(t)$ takes the form

$$h\left(t \mid \vec{Z}\right) = h_0(t) \exp\left(\vec{\beta} \cdot \vec{Z}\right), \tag{4.3}$$

where $h_0(t)$ is the baseline hazard, $\vec{Z}$ is the vector of covariates, and $\vec{\beta}$ is the vector of regression coefficients (to be estimated). The hazard is interpreted as the instantaneous probability of the event of interest occurring, given that it has not yet occurred, and ratios of hazard functions from either a known categorical contrast or a known increment in some continuous covariate are referred to as "hazard ratios," or HR. The maximum likelihood estimates (MLE) of $\vec{\beta}$ usually are accompanied by estimates of the standard error of the estimates for $\vec{\beta}$, and the significance of a given coefficient (i.e., β different from zero) is therefore reported via a Z-test. When a covariate takes on multiple but discrete "levels," (i.e., male versus female, over age 60 versus under age 60), the significance of the Cox PH model coefficient will closely match the significance of the equivalent log-rank test performed on the Kaplan-Meier curves for the discrete groups [63, 65]. This is due to the nature of the log-rank test, which is powered for and best suited for testing situations where the hazard ratio between groups is constant with time (the standard PH assumption).

The incorporation of time-varying covariates into a survival model is not trivial, since the fitting form of the Cox PH model changes. The partial likelihood function for the model becomes

$$L_P = \prod_{failure times\ j} \left(\frac{\exp\left(\vec{\beta} \cdot \vec{Z}_j\left(t_j\right)\right)}{\sum_{i \in R_j} \exp\left(\vec{\beta} \cdot \vec{Z}_i\left(t_j\right)\right)} \right), \tag{4.4}$$

where now the covariate vector $\vec{Z}(t)$ can change with time. The term R_j represents the patients

still at risk at the j^{th} failure time (i.e., still alive and on-study), and in a typical Cox PH model with fixed covariates, risk sets at later time points are subsets of earlier risk sets since patients can only leave the risk set, not re-enter. For patients who endure from one risk set to the next, their $\vec{\beta} \cdot \vec{Z}_i$ value does not change. In equation 4.4, though, there is additional computational complexity since the risk set R_j must be entirely recalculated at each unique failure time to properly account for covariates that change over time. While this extension of the PH model is not overly common, it is readily available in many statistical software packages [67].

Survival models can include examples of both time-invariant and time-varying covariates. For instance, the amount of disease can be allowed to vary over time, but the baseline covariates can remain fixed (such as the disease histology). By including multiple covariates in a survival model, one is said to be "adjusting" for other factors (e.g., the impact of changing disease volume among patients with a common disease histology). One could also include interactions between terms, such as allowing the coefficient determining the impact of changing disease burden to take on different values depending on baseline histology. All these capabilities are available in the academic edition of Revolution R Enterprise (version 4.3, based on R version 2.12; Revolution Analytics, Palo Alto, CA) [86].

In this study, patient survival is measured from a landmark time of the baseline CT scan prior to the initiation of treatment until either patient death or administrative censoring. No patients were lost to follow-up in this study, but a small number of patients remained living at the time of final database update in November of 2011. Survival needed to be defined from baseline rather than from diagnosis in order to have tumor volume data at time "zero," and while some patients in the study have CT scans taken during this pre-treatment or "natural history" time period, the processing and analysis of these natural history scans was not part of this dissertation. Clinical covariates available for each patient include sex, histology, weight, height, body surface area (BSA), TNM (tumor-node-metastasis) staging, International Mesothelioma Interest Group (IMIG) staging, smoking status, asbestos exposure, presence of chest pain, presence of dyspnea (shortness of

breath), weight loss noted prior to diagnosis, Eastern Cooperative Oncology Group (ECOG) performance status, talc pleurodesis status before chemotherapy, blood hemoglobin levels, white cell count, platelet count, baseline forced expiratory volume in one second (FEV1), baseline forced vital capacity (FVC), and age at diagnosis. All these covariates were fixed at their baseline values in the survival models. In concordance with prior prognostic models for mesothelioma, some covariates were binarized as follows: ECOG performance status was split to level 0 versus level 1 or 2, histology was split to epithelioid versus other, N staging was split to N0 versus other, IMIG staging was split to stage 1 or 2 or 3 versus stage 4, blood platelet count was split to ≤ 400 versus > 400 per nanoliter, blood hemoglobin was split to ≤ 140 versus > 140 grams per liter, and blood white cell count was split to ≤ 8.3 versus > 8.3 per nanoliter (these discretizations are found in [33] and [34]).

Tumor response measurements were allowed to change over time. Pleural disease volume measurements were modeled using scaled logarithmic transforms of relative changes from baseline, known as the specific growth rate (SGR) [106]. The definition of the SGR metric is

$$SGR(t) = \frac{\ln\left[\frac{V(t)}{V(t_0)}\right]}{(t - t_0)}, \tag{4.5}$$

where $V(t)$ denotes the volume measurement over time and t_0 indicates the time of baseline scanning (times in this study were all modeled as fractional years). The logarithmic transform of relative change data works well, since SGR is equal to zero for stable disease volumes, SGR is positive for growing disease volumes, and SGR is negative for shrinking disease volumes. Furthermore, the normalization by time span since baseline is useful since the time duration between patient CT scans was not strictly controlled in this study.

Survival models were constructed for this study using a forward selection process [88]. First, each covariate was tested one at a time in a Cox PH model on the full cohort of applicable patients, and covariates were noted for inclusion in the next step if their significance reached the

$\alpha = 0.10$ level. Next, the subset of covariates from the previous step were added one at a time to a multivariate Cox PH model. The covariates were added one at a time (starting with the most significant covariate from the previous step) until no further single addition achieved significance at the $\alpha = 0.10$ level. For the final multivariate model, covariates were selected from the previous step and added to the final model one at a time (in order of significance from the previous step) until no single addition resulted in a decrease of overall model AIC (Akaike's Information Criterion). AIC is defined as

$$\text{AIC} = 2k - 2\ln(L_P), \tag{4.6}$$

where k is the number of parameters in the model and L_P is the likelihood described in equation 4.4 [117]. AIC is a useful metric for assessing model fit since it seeks to improve overall fit through the likelihood term, but also avoids model over-fitting by penalizing the number of parameters in the model (for an example of AIC used for survival modeling with Cox PH models, see [118]).

While metrics like AIC give a *relative* sense of model performance (i.e., which model is the better of two options on the same group of patients), they do not give an absolute sense of model performance and are not applicable for models fit on different patient cohorts. Therefore, in order to assess the performance of a survival model, we need an appropriate metric. Receiver operating characteristic (ROC) analysis provides an interesting framework to assess survival curves. ROC analysis can be used when the data consist of an ordinal-scale predictor for a binomial outcome of interest and the true status of that outcome. Then, the sensitivity and specificity for predicting the outcome of interest can be calculated for a whole range of possible cut points in our predictor [68–70]. Plotting these sensitivity/specificity pairs on the typical ROC axes and calculating the area under the curve (AUC) gives one metric for how well the predictor performs.

In survival analysis models, at each time t, each patient will have a model-predicted probability of survival to time t, and of course each patient will have survived to time t or will have experienced the event of interest by time t. Therefore, treating our group of model-predicted survival probabilities for our patients as our set of "predictors," we can come up with an ROC curve for our model

at each time where a prediction is made by the model. For models with covariates fixed across time, Harrell's C statistic is equivalent to the non-parametric area under the fixed ROC curve [71]. However, if the covariates are allowed to vary with time, the rank ordering of predicted survival probabilities amongst patients can vary as well, and there will exist a unique ROC curve for model performance at each time. Each of these ROC curves will have a corresponding AUC, and plotting this AUC as a function of time gives us a general sense of our model performance. This is precisely the motivation for Heagerty's time-dependent AUC and C^τ [72], where we calculate our final performance metric as

$$C^\tau = \int_0^\tau \mathrm{AUC}(t) \cdot w^\tau(t)\, \mathrm{d}t, \tag{4.7}$$

where $\mathrm{AUC}(t)$ is the AUC as a function of time explained above, τ is the follow-up period of interest, and w^τ is a weight function. The problem has been considered by others [73,74], but Heagerty's metric is available in R. Although both Heagerty's C^τ and Harrell's C are identically scaled and in some sense should be interpreted similarly, C^τ and C can handle ties slightly differently and therefore their values are not directly comparable for situations with very coarse discretization in the predictor (as in Chapter 2, where there were only three possible classes).

To obtain estimates of predictive performance applicable to new cases, a cross-validation system is necessary [88]. The value of C^τ derived from both training and testing the model performance on the full cohort of applicable patients will be optimistic, and two methods were used in this study to cross validate model performance. First, a leave-one-out cross-validation (LOOCV) was used. Similarly to section 2.2.4, each of the 81 patients was omitted one at a time. A multivariate Cox PH model that consisted of the terms included in the final model as described above was fit to the remaining 80 patients, and the fit model was applied to the 81st patient. The predicted survival over time was stored for this single omitted patient. The process was repeated until each of the 81 patients had been excluded once from the training cohort. In this way, predicted survival for all 81 patients was calculated using models that had no knowledge of the patient in question. A value of C^τ was calculated from these LOOCV-based survival predictions, called C^τ_{cv}, which

is a better indicator of how well the reported final Cox model will perform for a new, previously unknown patient.

The second validation procedure was performed via repeated random sub-sampling and was intended to assess the impact of any small number of patients on the overall model performance as well as generate a confidence interval for model performance (standard errors of C^{τ} are not readily reported using Heagerty's package in R). For each sub-sampling iteration, a model was trained on two-thirds of the patient cohort and tested on the remaining one-third of the patient cohort. The training set was chosen randomly without replacement at each iteration, and the testing set was considered to be the remaining patients who had not been selected for the training set at that iteration (there are over 10^{21} possible unique combinations of 81 patients into a training set of 54 and a testing set of 27). After fitting the final multivariate Cox PH survival model described above to the training cohort, the value of C^{τ} from the testing set was tabulated. The repeated random sub-sampling simulation consisted of 1000 iterations, and the mean performance metric value will be referred to as C^{τ}_{sub}. The sub-sample performance 95% confidence interval was calculated from the 2.5% and 97.5% quantiles of the tabulated sub-sample performance scores.

4.3 Results

4.3.1 Patients and Overall Survival

Median survival from baseline imaging prior to treatment initiation was 12.6 months (95% confidence interval, 10.6–14.9 months). Of the 81 patients, there were 76 observed deaths, while the remaining five patients were lost to follow-up after a median duration of 35 months. The overall Kaplan-Meier survival curve is shown in Figure 4.8.

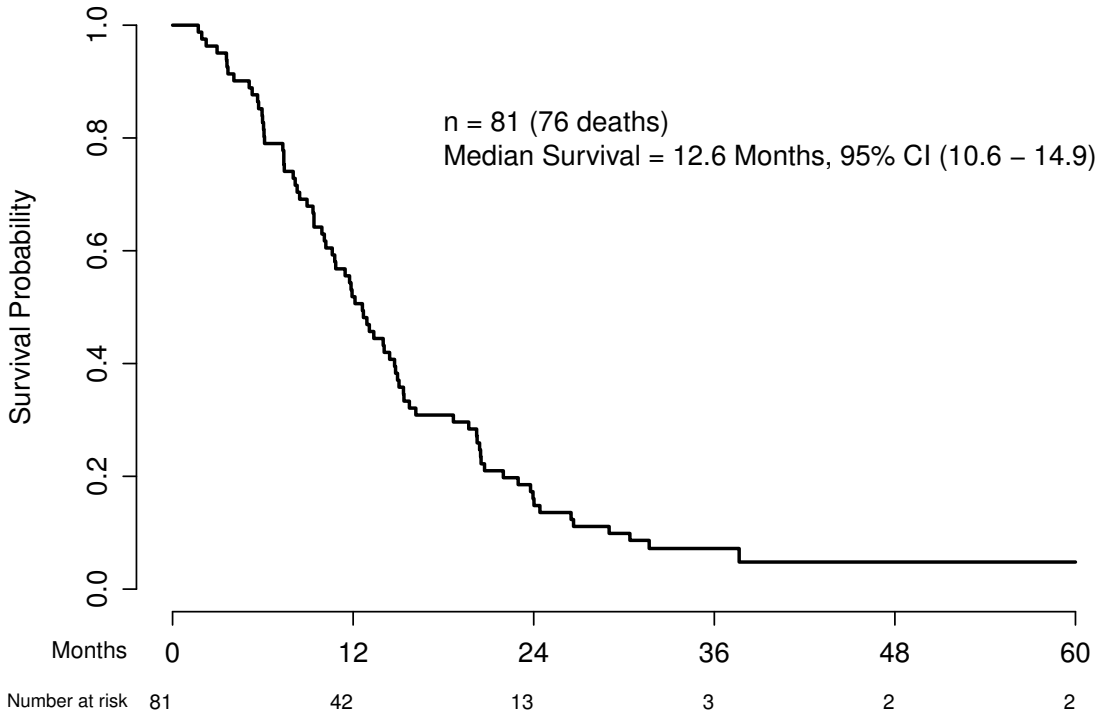


Figure 4.8: Overall survival for the patient cohort in this study.

4.3.2 Disease Segmentation

Across all patients, the mean pleural disease volume at baseline was 1511 ± 1065 mL (range 225–5287 mL). For patients with stage 1, 2, or 3 disease, the mean disease volume at baseline was 1394 mL, whereas the mean for stage 4 patients was 1744 mL (the difference was not significant by a Student's t-test, $p = 0.17$). At first follow-up, the mean disease volume had reduced to 1397 mL, with geometric mean change from baseline of -11%. By the end of treatment, the geometric mean change in disease volume from baseline was -19%.

A threshold was determined to split patients into two groups (large disease volume and small disease volume) using a minimum p-value technique, after which the p-value was automatically adjusted to correct for multiple hypothesis testing [119, 120]. The cut-point was determined to be a baseline disease volume of 522.3 mL, and the resulting large disease volume cohort (n=68) had a median survival of 11.9 months, while the small disease volume cohort (n=13) had a median survival of 19.7 months. However, after appropriate adjustment of the p-value, the difference was not significant ($p = 0.32$). The Kaplan-Meier survival curves for both groups are shown in Figure 4.9.

Because the disease segmentation process required manual intervention for every individual CT scan, the intent was to perform disease segmentation only on a single reconstruction series for each unique patient scan. However, segmentation was unintentionally performed on two distinct axial reconstructions of differing slice thicknesses for five patient scans. These five cases can provide some minimal level of variability quantification for disease segmentation volumes derived from a single observer. The median unsigned difference between the estimates of disease volume was 4.2% (for absolute differences, the disease volume from the reconstruction with thinner axial slices was 3.0% larger on average).

To quantify the poor agreement between the automated disease segmentation algorithm using the attempted rib segmentation and the disease segmentation from semi-automated hemithoracic segmentations, the Jaccard similarity coefficient had a median value of $J = 0.27$ (without

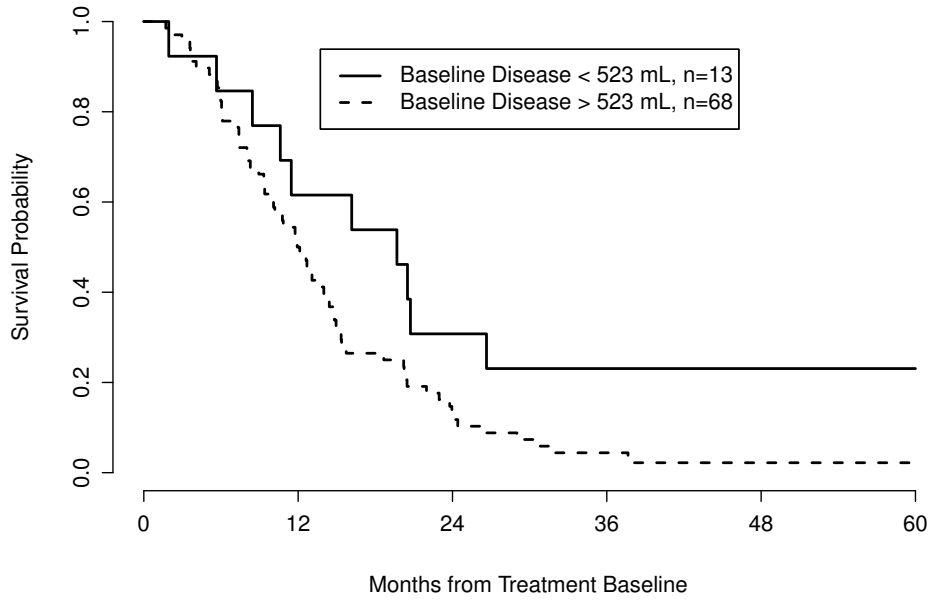


Figure 4.9: Survival curves for patients with baseline disease volumes above and below 522.3 mL.

the attempted rib segmentation, the similarity between automated and semi-automated disease segmentations was even worse, with a median value of $J = 0.25$). Following the initial disease segmentation from the semi-automated hemithoracic segmentations, the disease segmentations were all manually corrected, though this correction was minimal, as evidenced by a median value of $J = 0.97$ between semi-automated disease segmentations and manually corrected disease segmentations. As mentioned in section 4.2.3.2, the comparison between hemithoracic segmentation approaches would be misleading, since their aims were to contour inherently different anatomic structures and therefore those results are not given here.

Segmentations were all performed on a 64-bit PC running the latest version of MATLAB at the time of computation (currently version 7.13), with 8 GB of RAM and a quad-core processor with a clock speed of 2.66 GHz. Much of the code is not optimized for multi-core computation, but the timings given below are perhaps useful for reference. The automated thoracic segmentation required 2.0 minutes on average and the bone and contrast segmentation required 2.8 minutes on

average, which both scale approximately linearly with the number of axial sections. The automated airway segmentation required 27.2 minutes on average, and the automated lung segmentation required 14.7 minutes on average. While these automated steps together require 46.7 minutes on average, it should be pointed out again that these are 46.7 minutes of purely background computation: the user can start the segmentation process for hundreds of scans at a time and simply walk away (or leave on vacation). Additionally, these algorithms could potentially be sped-up by a factor of two to three with already existing parallelization routines in MATLAB.

For completeness, the timings for the remaining automated segmentation algorithms are given here, although the semi-automated techniques were used to compute the volumes in this study. Prior to hemithoracic segmentation, the ribcage segmentation required 7.0 minutes on average. Non-linear diffusion filtering of the image volume required 14.3 minutes on average, and the final hemithoracic boundary segmentation required 6.2 minutes on average. The pleural disease segmentation algorithm required 5.2 minutes on average, 2.4 of which were devoted to volumetric watershed region splitting.

For the semi-automated segmentation technique used in this study, there are two additional timings that are relevant. First, the time required to complete a semi-automated segmentation of the hemithoracic boundary from contour seeding through final interpolated contour approval is estimated to be between 10 and 15 minutes by the author, though no formal timings were performed. For cases with unilateral disease, especially with disease limited inferiorly by the diaphragm, segmentation may be completed in as little as five minutes, but for cases with bilateral disease or complicated disease boundaries, segmentation may require 15 minutes. Diffusion filtering of the image volume is still a required pre-processing step for the pleural disease segmentation, and the pleural disease segmentation itself is unchanged with regards to timing (only the hemithoracic input is changed). Finally, because the semi-automated hemithoracic segmentations are manually constrained to the outer border of disease, the manual correction process for the pleural disease segmentations is minimal, on average requiring less than five minutes (predominantly to clean up

spurious and spatially isolated labeled watershed volumes). The same manual correction process on the *automated* pleural disease segmentations typically required 45 minutes per case, from the author's limited experience in processing such automated segmentations (approximately only 10 cases were attempted in this way). Therefore, while the semi-automated approach used in this study requires perhaps 15–20 minutes of active user intervention per case on average, correcting the output of the fully automated approach required around 45 minutes of active user intervention per case.

4.3.3 Univariate Survival Analysis

Covariates that were predictive for survival in univariate Cox PH models at the $\alpha = 0.10$ level are shown in Table 4.2. Any covariates mentioned in section 4.2.4 but not shown in Table 4.2 had p-values larger than 0.10 and were therefore not included in future analyses. The clinical covariates predictive for survival are not novel discoveries, since all have been reported in previous prognostic models for patients with mesothelioma. However, continuous time-changing measurements of pleural disease volume (modeled according to equation 4.5) are significantly associated with patient survival, and this is a novel discovery. As would be expected, the larger the disease SGR, the larger the hazard ratio. Therefore, patients with disease growth have worse prognosis than patients with disease shrinkage, and patients with substantial growth have worse prognosis than patients with minimal growth. Kaplan-Meier curves comparing survival between the different levels of histology, dyspnea, ECOG performance status, and M stage are shown in Figures 4.10, 4.11, 4.12, and 4.13, respectively.

4.3.4 Multivariate Survival Analysis

In the first multivariate Cox PH model, covariates from Table 4.2 were added one at a time (starting with the most significant covariate from the previous step) until no further single addition achieved significance at the $\alpha = 0.10$ level. This resulted in a model including disease volume

Variable		Hazard Ratio	95% CI	p-value
N stage	N0	1	–	–
	N1+	1.58	[0.95, 2.61]	0.076
M Stage	M0	1	–	–
	M1	1.98	[1.03, 3.81]	0.040
Dyspnea	No	1	–	–
	Yes	1.94	[0.98, 3.82]	0.056
Weight Loss	(continuous, in kg)	1.05	[1.00, 1.10]	0.054
Talc Pleurodesis	No	1	–	–
	Yes	0.67	[0.41, 1.08]	0.098
ECOG Performance Status	0	1	–	–
	1 or 2	1.56	[0.98, 2.46]	0.059
Histology	Epithelioid	1	–	–
	Other	2.26	[1.34, 3.83]	0.0023
Blood Platelet Count	≤ 400 per nanoliter	1	–	–
	> 400 per nanoliter	1.58	[0.97, 2.57]	0.067
Disease Volume	(continuous, SGR)	1.26	[1.11, 1.42]	0.00031

Table 4.2: Factors predictive for survival (from baseline imaging) in univariate Cox PH models, including hazard ratios and 95% confidence intervals (CI). Disease volume modeled as continuous specific growth rate (SGR) from baseline.

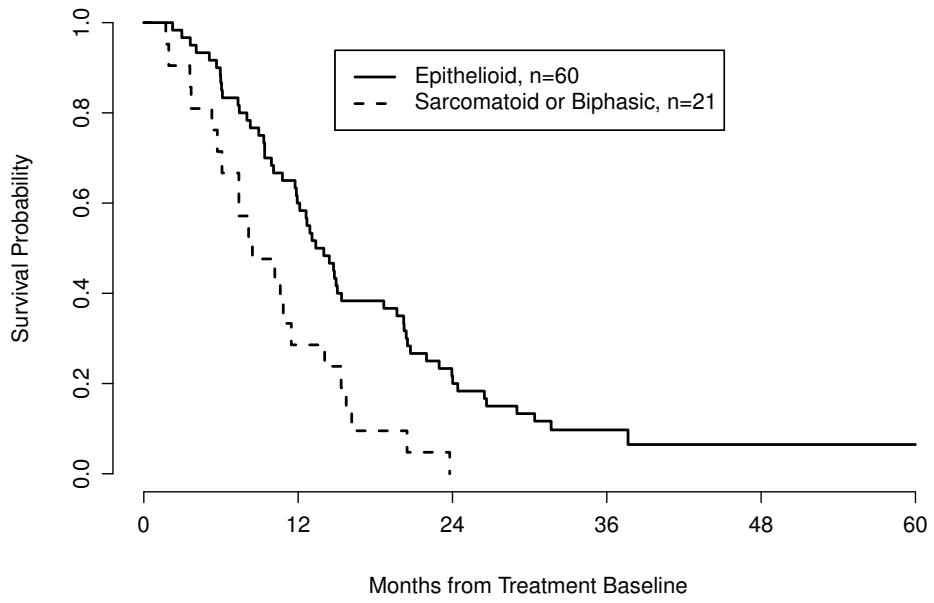


Figure 4.10: Survival curves for different values of the histology covariate (log-rank test $p = 0.0018$).

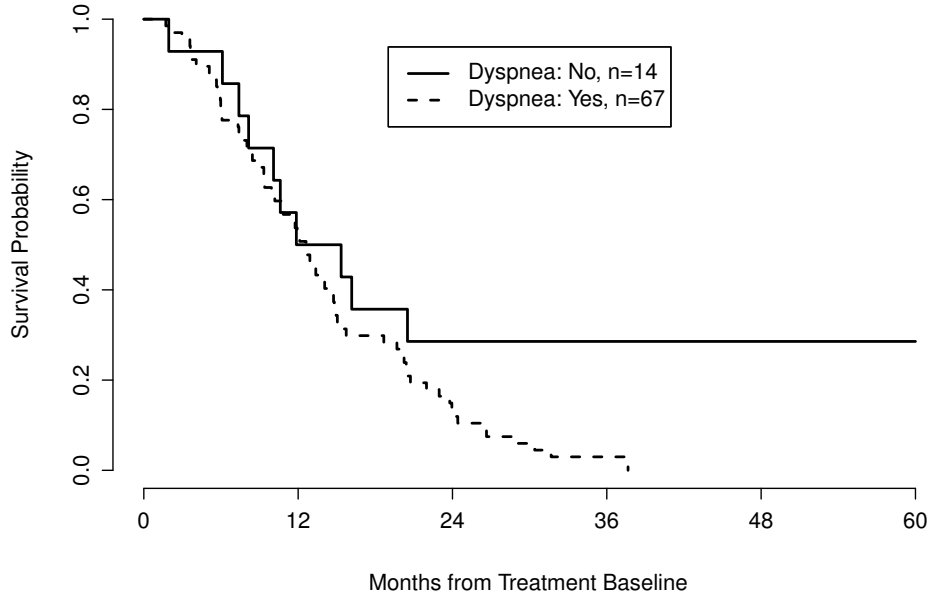


Figure 4.11: Survival curves for different values of the dyspnea covariate (log-rank test $p = 0.052$).

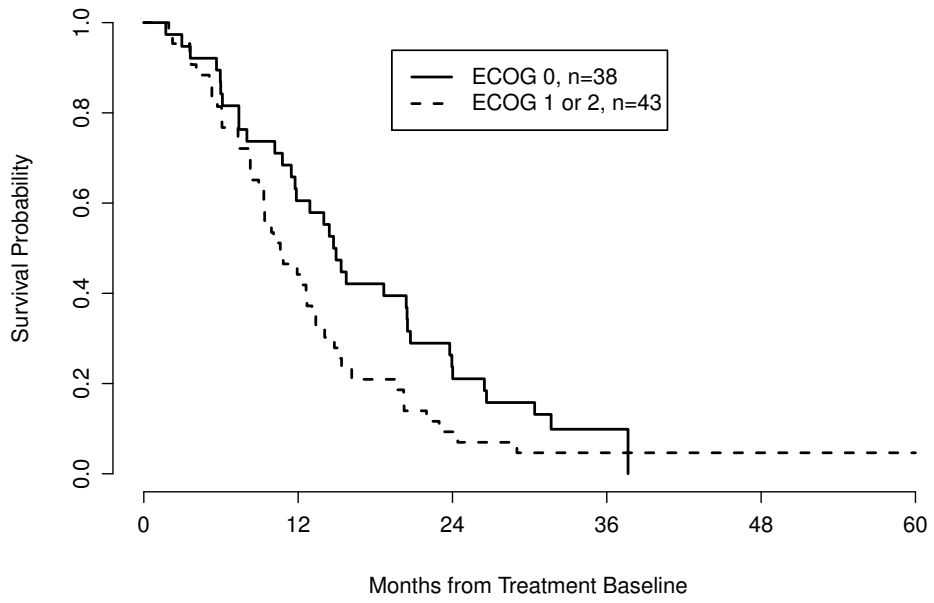


Figure 4.12: Survival curves for different values of the ECOG performance status covariate (log-rank test $p = 0.057$).

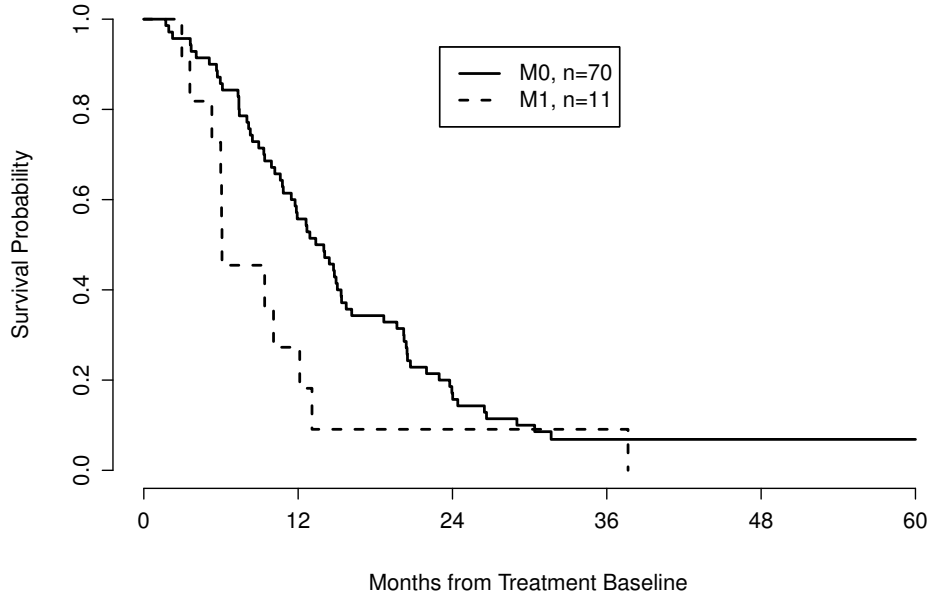


Figure 4.13: Survival curves for different values of the M stage covariate (log-rank test $p = 0.036$).

SGR (HR=1.30, $p = 0.00088$), histology (HR=2.33, $p = 0.0025$), M stage (HR=1.67, $p = 0.097$), dyspnea (HR=2.94, $p = 0.0044$), and ECOG performance status (HR=1.66, $p = 0.039$). The final multivariate model was constructed by adding these covariates one at a time in order of increasing p-value until no addition (or replacement) achieved a reduction in AIC. The AIC values of the successive models are shown in Table 4.3, leading to the final model described in Table 4.4 (model 5 from the AIC table). The included covariates are disease volume SGR (HR=1.31, $p = 0.00045$), histology (HR=2.28, $p = 0.0029$), dyspnea (HR=3.20, $p = 0.0020$), and ECOG performance status (HR=1.69, $p = 0.029$). No two-way interactions between covariates in this final multivariate model were significant.

The value of C^{τ} was calculated for the final multivariate model on the full patient cohort (after training on the full patient cohort). The value used for τ was the time associated with the final observed patient death (37.7 months), and the calculated performance value was $C^{\tau} = 0.690$. Because of differences in the calculation of Harrell's C (used in Chapter 2) and Heagerty's C^{τ}, values of C^{τ} are *not* directly comparable to values of C from Chapter 2; for a comparison of survival model performance for disease volumes and linear measurements, the reader is referred to the following chapter. From the leave-one-out cross-validation, the survival model performance from the final multivariate model was $C^{\tau}_{cv} = 0.661$, a slight reduction from the full cohort performance (as predicted). The 1000 random sub-sampling iterations yielded a mean model performance of $C^{\tau}_{sub} = 0.660$, and the 95% confidence interval of sub-sample performance values ranged from 0.556–0.746. Figure 4.14 shows the histogram of sub-sample performance values from the 1000 simulation iterations.

4.4 Discussion

The goal of this study was to create a survival model using time-varying image-based measurements of pleural disease volume modeled as "specific growth rate" from baseline for patients with malignant pleural mesothelioma. Disease volumes extracted from semi-automated disease seg-

Model Number	Model Covariates	AIC
1	None (Null Model)	542.09
2	Disease Volume SGR	531.55
3	Disease Volume SGR, Histology	528.23
4	Disease Volume SGR, Histology, Dyspnea	520.05
5	Disease Volume SGR, Histology, Dyspnea, ECOG Performance Status	517.26
6	Disease Volume SGR, Histology, Dyspnea, ECOG Performance Status, M Stage	517.30

Table 4.3: Values of Akaike's Information Criteria (AIC) for forward selection iterations of the final multivariate Cox PH model.

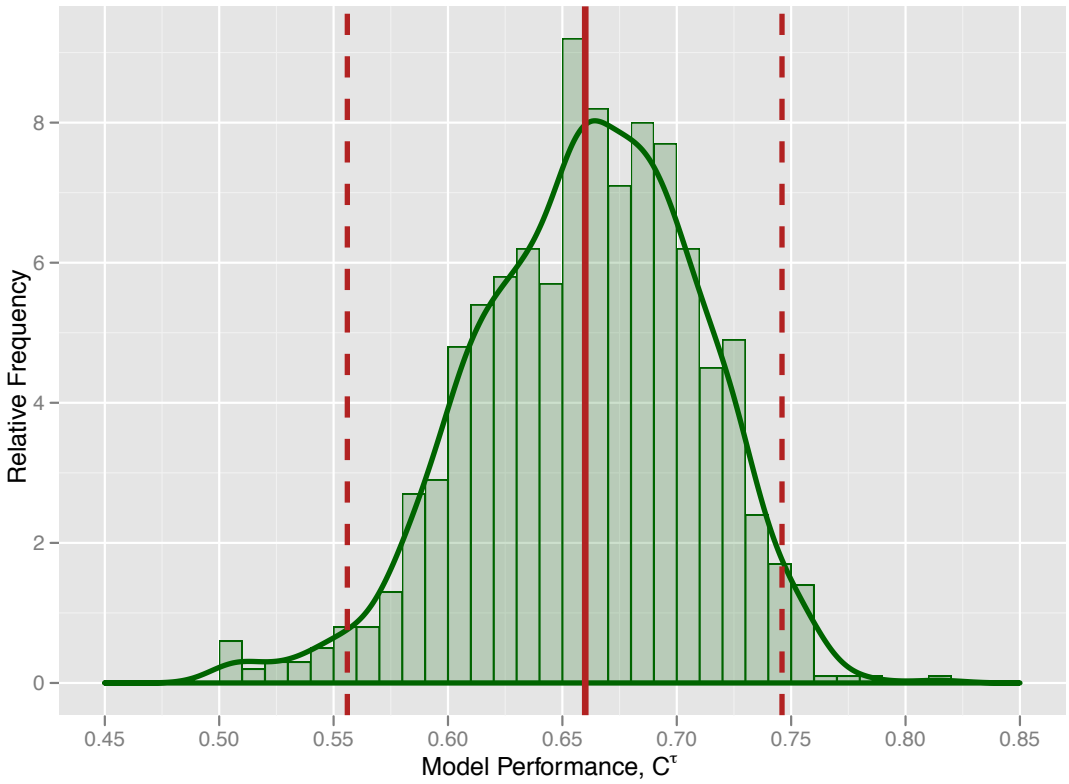


Figure 4.14: Smoothed histogram of repeated random sub-sample performance values C^{τ} for the final multivariate Cox PH model trained on two-thirds of the patient cohort selected at random and tested on the remaining third (1000 iterations). Mean and 95% confidence interval shown with red lines.

Variable		Hazard Ratio	95% CI	p-value
Disease Volume	(continuous, SGR)	1.311	[1.13, 1.53]	0.00045
Histology	Epithelioid	1	–	–
	Other	2.283	[1.33, 3.93]	0.0029
Dyspnea	No	1	–	–
	Yes	3.204	[1.53, 6.71]	0.0020
ECOG Performance Status	0	1	–	–
	1 or 2	1.693	[1.05, 2.72]	0.029

Table 4.4: Factors predictive for survival in the final multivariate Cox PH model, including hazard ratios and 95% confidence intervals (CI). Disease volume modeled as continuous specific growth rate (SGR) from baseline.

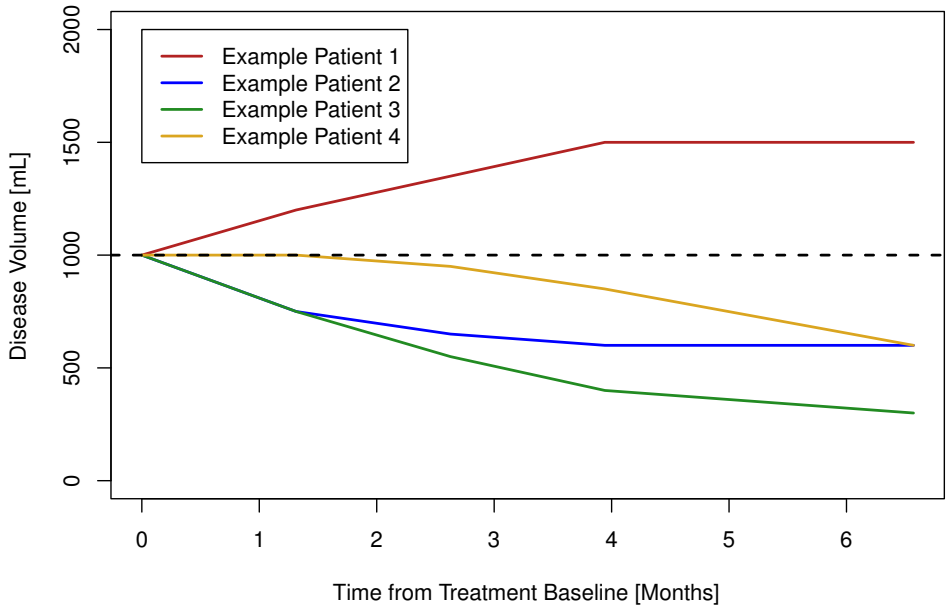
mentations proved to be significant predictors of patient survival in conjunction with other clinical covariates. In fact, the p-value associated with disease volume for survival prediction is the smallest of any of the included terms (in either the univariate or multivariate analysis), indicating the pronounced prognostic value of the image-based measurements for tumor response assessment.

Because of the continuous and time-varying nature of the specific growth rate (SGR) metric used to model changes in disease volume, interpretation of the Cox PH model terms can be slightly challenging. Because the hazard ratio is larger than unity, tumor growth (i.e., positive SGR) is associated with poor prognosis, and tumor shrinkage (i.e., negative SGR) is associated with more positive prognosis. Additionally, the larger the tumor growth (i.e., larger SGR), the worse the prognosis. This can be better illustrated using some visual examples. Figure 4.15a shows volume trajectories for four hypothetical patients, and Figure 4.15b shows the equivalent SGR trajectories for the same patients. Next, each patient was given the same set of clinical covariates (epithelioid histology, ECOG performance status of 0, and noted presence of dyspnea). Survival probability over time was predicted for each patient using the final multivariate model in Table 4.4, and the resulting survival curves are shown in Figure 4.16. Predicted survival probabilities at 12 months

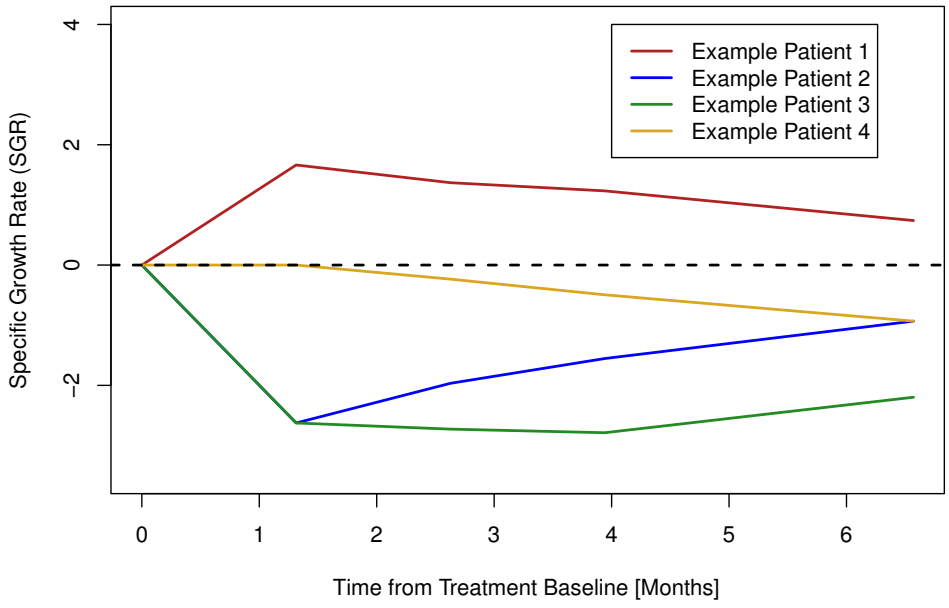
are 54%, 71%, 78%, and 68% for patients 1, 2, 3, and 4, respectively, and at 24 months the survival probabilities are 11%, 25%, 37%, and 24%, respectively.

Some of the unique contributions of this study are the ability to model changes in disease volume through time and the modeling of continuous (not discretized) changes in disease volume. For instance, in the only other study using *changes* in disease volume as a prognostic covariate for patients with MPM by Liu *et al*. [58], analysis was limited to a single follow-up point and changes in disease volume were discretized as "volume decrease" or "volume increase." In the current work, any arbitrary number of follow-up points can be used, and changes in disease volume are treated continuously. In the work by Liu *et al*. the hypothetical patients #2 and #3 in Figure 4.15 would be interpreted identically after the first follow-up volume quantification, and subsequent updates to the volume trajectory would be ignored. In this study, however, the survival curves for patients #2 and #3 clearly separate in Figure 4.16 after more of the disease volume trajectory has been realized. In the patient cohort of this study, fitting a model similar to Liu *et al*. where the only prognostic covariate was discrete volume change after first follow-up yielded a model with a C^{τ} value of 0.521 (compare to $C^{\tau} = 0.690$ for the full multivariate model in this study).

Although this study is not the first to use changes in MPM disease volume as a prognostic marker [37, 54–58], it is the first to model changing disease volumes using the SGR metric, and is also the first (to the author's knowledge) to combine clinical covariates with changing disease volumes in a multivariate survival model. This is an important distinction, since it was shown in this work that the addition of baseline clinical covariates to the multivariate model results in a reduction of AIC. Although it was not shown in the results above, the C^{τ} value for a model *only* including the disease volume term was 0.610, while for the final multivariate model also including the baseline covariates $C^{\tau} = 0.690$. Finally, just for comparison, a multivariate model built using the forward selection technique described above using only baseline clinical covariates (i.e., no disease volumes) resulted in a model including histology, dyspnea, ECOG performance status, and M stage, and the full cohort C^{τ} value was 0.659. Therefore disease volume changes and the clinical



(a)



(b)

Figure 4.15: 4.15a, disease volume trajectories for four hypothetical patients. 4.15b, corresponding specific growth rate trajectories for the same four patients (time in equation 4.5 adjusted to years).

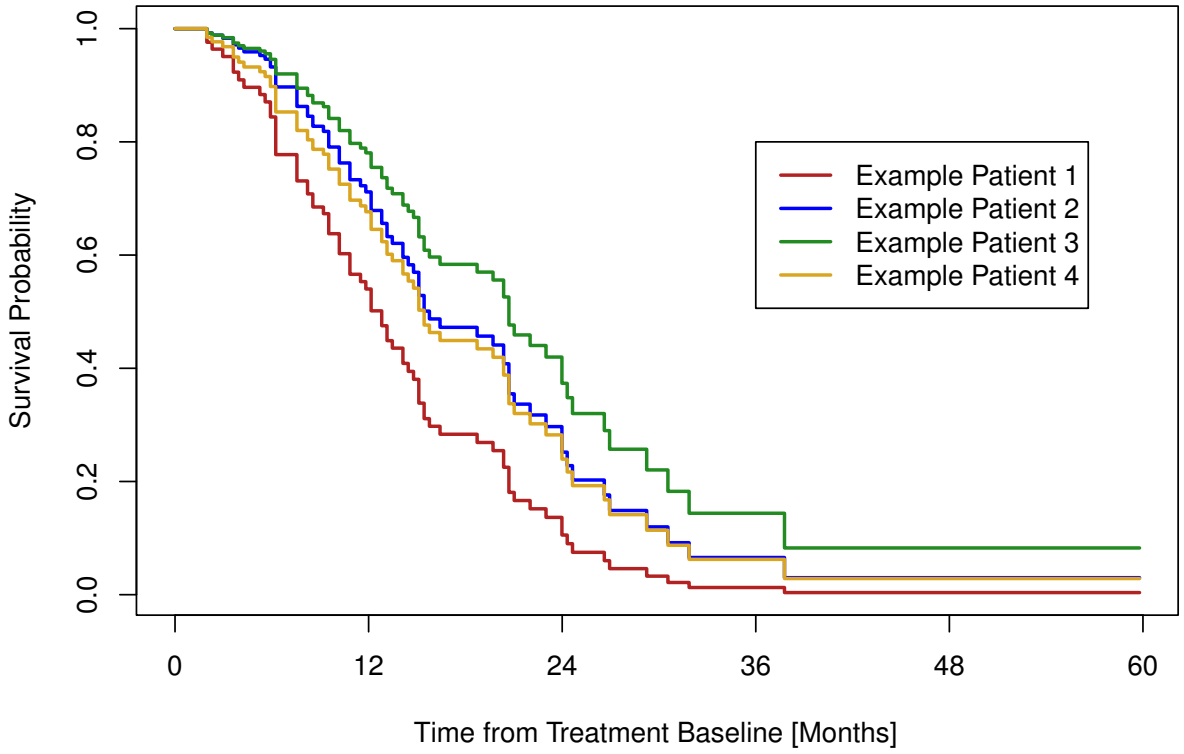


Figure 4.16: Predicted survival curves for the four hypothetical patients whose volume trajectories are shown in Figure 4.15, using the survival model in Table 4.4. All patients were assigned epithelioid histology, ECOG performance status of 0, and presence of dyspnea.

covariates are both crucial components of the final multivariate survival model.

The modeling of disease volume SGR as a continuous and time-varying quantity has some clear advantages. First, in any attempt to discretize disease volume changes into two or more classes, the class distinctions will be arbitrary. While one cut-point for changes is at 0% (i.e., growth versus shrinkage) and this cut-point has been used in previous studies with MPM disease volumes on CT [58], changes in disease volume are inherently continuous. In Frauenfelder *et al.* [57], for example, the authors use an extrapolation of the RECIST classification criteria to volumetric changes based on a spherical geometry, which remain entirely arbitrary for reasons expounded upon elsewhere in this dissertation. The modeling of changes in disease volume as time-varying is a bit of a double-edged sword: the model fitting and interpretation is more complicated, but flexible and arbitrary volumetric trajectories can be modeled for each patient. In many published studies, there are exactly two volume measurements for each patient, baseline and "follow-up." Follow-up can be defined after a set number of chemotherapy cycles, but what happens before or after that milestone? If a patient gets four follow-up scans during treatment, the researchers have updated information about how the patient's disease volume has changed since baseline. The modeling approach used in this study allows the researchers to utilize the updated information to predict patient survival, whereas other fixed follow-up models cannot take advantage of the updated knowledge. Figures 4.15 and 4.16 highlight the substantial impact on predicted survival that updates in patient disease volume can have.

One disadvantage of the disease segmentation technique used in this study is that segmenting pleural disease is not necessarily the same as segmenting mesothelioma tumor *per se*. For cases in which pleural effusion is present and spatially mixed with tumor tissue, the separation of tumor from effusion can be difficult to impossible. The same challenge was faced by Sensakovic *et al.* [60] in the creation of an automated segmentation technique. The only difference in the segmentation scheme of this study and that of Sensakovic *et al.* is the semi-automatic hemithorax segmentation. The pleural space segmentation component is identical between the two studies

and therefore the automated handling of pleural effusion is limited by the ability of the k-means classifier to separate tissue types based on HU values. In this study, pleural effusion regions were manually excluded if they were localized and spatially independent of tumor tissue regions, but when effusion was mixed together with tumor tissue, no attempt was made to exclude the effusion. It is unknown whether quantification of true tumor volume has more prognostic value than pleural disease volume. Certainly pleural effusion and tumor tissue both have negative impact on patient quality of life, but clinical management is different for the two types of pleural disease, and future work could potentially focus on the construction of a prognostic model from time-varying measurements of purely tumor tissue.

Although the task of pleural disease segmentation is greatly facilitated through as much automation as possible, it will always be important for a trained observer to "proof" the final segmentations and manually correct the results. It is not unreasonable to fear that this manual process may introduce undesirable imprecision in the final volumetric measurement. However, the results of Frauenfelder *et al.* [57] indicate that "volumetry is highly reliable, reproducible and reader independent compared with the modified RECIST criteria." The comparison in that study was made using discretizations of changes in disease volume according to volumetric response criteria extrapolated from the spherical geometry described above, where the cutoffs for PR/PD were -65%/+73%. Therefore, the increased inter-observer agreement for volumetric measurements over linear measurements is likely in part due to the widening of the stable disease category alone (the standard RECIST criteria of -30%/+20% were used for linear measurements). However, the general conclusion is still important and pertinent to the current study. The very small number of cases in this study for which two reconstruction series were segmented for the same scan also point to the robustness of volumetric quantification by a single observer.

The inclusion of disease histology, dyspnea, and ECOG performance status in the final multivariate model is well in line with previous prognostic models for MPM (see especially reference [31]). The indication that non-epithelioid histology, presence of dyspnea, and non-zero ECOG

performance status are associated with poor prognosis is well-known for MPM patients. These factors are routinely available as part of the diagnostic workup of patients with MPM, and they improve the predictive performance of the resulting survival model. Theoretically, the presence of dyspnea and ECOG performance status can and do change over time for each patient and could therefore be treated as time-varying in the multivariate prognostic model. However, these data are not typically recorded or available after the baseline assessment of each patient, and therefore the treatment of these covariates as fixed in time is in line with current clinical practice.

It is important to acknowledge this study's limited patient cohort size. Increasing the patient cohort may reveal associations between new covariates and survival, potentially improving the performance of the survival model. We hope to validate the model performance in a larger independent patient cohort. Unfortunately, such cohorts are difficult to obtain or construct, especially cohorts with serial CT imaging and a similar treatment regimen. Even after the collation of a patient database, pleural disease segmentations are time-consuming and tedious. The observer tasked with semi-automated hemithorax segmentation in this study had intimate knowledge of the segmentation technique, and a duration of 10–15 minutes per case was only achievable after familiarization and practice (compare this with the approximately two hours required for a complete manual disease segmentation). The feasibility of semi-automated hemithoracic segmentation and subsequent disease segmentation editing in the clinical workflow is unclear at this point, and therefore the routine clinical use of disease volumes for MPM prognosis remains to be investigated. Finally, the patients in this study were treated with cytotoxic chemotherapy, not with other therapies such as anti-angiogenic therapy, but treatment-specific models could be developed for individual patient cohorts. The model derived in this study may not be directly applicable to patient cohorts receiving biologically different treatments.

In summary, volumes derived from semi-automatic segmentations of pleural disease in patients with MPM are prognostically significant. Disease volume (specifically the "specific growth rate") is allowed to change over time, allowing for survival prediction from arbitrary disease volume

trajectories. In a final multivariate model, disease histology, dyspnea, and ECOG performance status were also significant predictors of patient survival.

CHAPTER 5

MEASURING PATIENT RESPONSE: ALTERNATIVES TO DISEASE VOLUME

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny…'" – Isaac Asimov

5.1 Introduction

For matters involving tumor response, there is only one metric that can be considered to be "truth": the proliferative cellular tumor burden. If we can convince ourselves that the tumor bulk is spatially homogeneous in cellular density, changes in tumor volume will directly correspond to changes in the number of tumor cells. Some molecular imaging methods, such as total glycolytic volume (TGV) quantification in fluorodeoxyglucose-positron emission tomography (FDG-PET), are moving toward proliferative cellular quantification. However, until these methods become widespread, computed tomography (CT) imaging with the possibility of volumetric quantification will remain the best tool to assess patient tumor burden for patients with malignant pleural mesothelioma (MPM).

As described in Chapter 4, mesothelioma segmentation and volumetric quantification is a challenging task. The morphology of the disease is widely variable between patients, and the radiographic density of the disease mimics tissues often in close proximity to the tumor itself. Some degree of manual intervention is necessary for segmentation of MPM in CT scans, a task potentially requiring up to 20 minutes per scan. The disease volume segmentation process is further complicated by the occasional presence of pleural effusion, the identification and exclusion of which can be difficult to impossible. While volume measurements of MPM have been shown to exhibit lower inter-observer variability than linear thickness measurements made according to the modified Response Evaluation Criteria In Solid Tumors (RECIST) protocol [57], the computational and

manual challenges of the volume segmentation task are problematic.

Since we cannot count individual tumor cells, it is necessary to quantify changes in tumor burden using medical images. It was previously shown that pleural disease volume is a significant predictor of patient survival, but there are other ways to answer the question "is the tumor burden growing or shrinking?" Chapter 1, specifically section 1.2.2, argued that the current application of the modified RECIST measurement technique with the standard RECIST classification criteria is not specifically suited to mesothelioma. Although Chapter 2 was devoted to the optimization of response classification criteria for linear thickness measurements, there is reason to believe that continuous (not discretized) changes in tumor burden should be used for response assessment [44]. The modified RECIST measurement technique may be an oversimplified tool to measure tumor burden, but it is a tool specifically suited for the morphology of MPM. By focusing on continuous changes in tumor burden as measured using mesothelioma-specific linear thickness measurements (potentially in conjunction with other clinical covariates), it may be possible to improve the association between measured tumor response and actual patient survival.

Because of the anatomy associated with MPM, changing tumor burden affects more than just the volume of tumor. The hemithoracic space is fairly fixed by the ribcage and mediastinal anatomy. Therefore, when the disease volume increases, it is reasonable to believe that the aerated lung volume may decrease correspondingly. The disease volume and aerated lung volume are physiologically correlated, and while the trade-off will very likely *not* be one-to-one, changes in lung volume may have prognostic value for patients with MPM. Lung volume has been used to monitor response in MPM patients previously, though the setting was after surgical intervention where lung re-expansion was the expected result after comprehensive tumor de-bulking [108]. Changes in lung volume may also be a useful tool to assess tumor response for patients receiving chemotherapy; instead of classifying response from declining tumor volume, response would be classified from increasing lung volume.

Both linear measurements and lung volumes have advantages over disease volumes for re-

sponse assessment. Disease volumes require substantial manual intervention. Linear thickness measurements are almost entirely manual (though some automation techniques have been suggested [121]) but require much less time than disease volume segmentation. Lung volume segmentation, on the other hand, is entirely automated and is likely to be the measurement technique with the lowest variability of the three techniques. The purpose of this study is to build prognostic models for patients with MPM using changing lung volumes or changing linear thickness measurements in place of changing disease volumes. The performance of these models will be compared with the models built in the previous chapter. The hypothesis of this study is that lung volumes and linear measurements, treated continuously, will both be prognostic for patient survival for MPM patients during chemotherapy.

5.2 Patients and Methods

5.2.1 Patient Cohort

The patient cohort used in this study was a subset of the entire "Perth database" described in section 2.2.1 and Table 2.1. For inclusion in this specific study, patients were required to have available modified RECIST tumor thickness measurements at baseline (prior to beginning chemotherapy) and for one or more follow-up scans during chemotherapy. The patients were also required to have unilateral disease, since the analysis below for lung volumes is limited to patients with one non-diseased lung. Finally, all patients were required to have a complete CT scan for all scan dates (i.e., not simply scanned films) for automated lung segmentation. These constraints reduced the eligible patient cohort to 61 patients, and the summary description of these patients is given in Table 5.1.

Table 5.1: Description of the patient cohort used in this specific study, consisting of 61 of the original 97 patients. This specific patient cohort is a subset of the patients summarized in Table 2.1.

Characteristic	Summary
Sex:	
Male	n = 50
Female	n = 11
Age at Diagnosis:	
Median	66 years
Range	42–80 years
Chemotherapy:	
Carboplatin/Pemetrexed	n = 6
Cisplatin/Pemetrexed	n = 31
Cisplatin/Gemcitabine	n = 24
Histology:	
Epithelioid	n = 43
Sarcomatoid	n = 5
Biphasic	n = 13
T Stage:	
T1	n = 13
T2	n = 16
T3	n = 20
T4	n = 12

(continued on next page)

(Table 5.1, continued from previous page)

N Stage:		
	N0	n = 17
	N1	n = 2
	N2	n = 32
	N3	n = 10
M Stage:		
	M0	n = 55
	M1	n = 6
IMIG Stage:		
	I	n = 9
	II	n = 2
	III	n = 29
	IV	n = 21
Known Asbestos Exposure:		
	Yes	n = 55
	No	n = 6
Chest Pain:		
	Yes	n = 38
	No	n = 23
Shortness of Breath:		
	Yes	n = 50
	No	n = 11

(continued on next page)

(Table 5.1, continued from previous page)

ECOG Performance Status:		
	0	n = 31
	1	n = 26
	2	n = 4
Talc Pleurodesis:		
	Yes	n = 27
	No	n = 34
Weight:		
	Median	75 kg
	Range	52–121 kg
Height:		
	Median	171 cm
	Range	155–188 cm
Smoking Status:		
	Never	n = 27
	Past	n = 29
	Present	n = 5
Pleurectomy/Decortication:		
	Yes	n = 0
	No	n = 61

5.2.2 Imaging

Patients were imaged using helical CT up to one month prior to the first cycle of chemotherapy and throughout their treatment regimen (typically after the first cycle, then every two cycles there-

after). CT staging was performed according to the Union for International Cancer Control (UICC) TNM staging system (2002). CT scans were staged by a thoracic radiologist or medical oncologist experienced in mesothelioma imaging. Tumor thickness measurements were made on 5 mm reconstructions according to the modified RECIST protocol on baseline and all follow-up scans [47].

There were a total of 216 CT scans in this study, with a median of four scans per patient (including baseline scans). Eight patients had only a baseline scan with one follow-up scan, while 19 patients had three scans total, 27 patients had four scans total, and seven patients had five scans total. The median duration between scans was 48 days. Of the 216 scans, 150 were performed on General Electric scanners (HiSpeed CT/i, n=81; LightSpeed Pro 16, n=1; or LightSpeed VCT, n=68), and 66 were performed on Philips Brilliance 64-slice scanners. At least 101 of the scans were performed with iodinated contrast media (for the other 115 scans, the contrast field in the DICOM image header is empty, which does not necessarily imply that no contrast was administered).

Only one reconstructed series was required for segmentation for each CT scan date, and this series was selected for each patient with consideration for reconstruction kernel and slice thickness. The decisions regarding the choice of reconstructed series for each scan were made for the disease volume segmentation task described in Chapter 4. Although linear thickness measurements were consistently acquired using 5 mm reconstructions, multiple reconstructed slice thicknesses exist for each CT scan. For the scans used in this study, slice thickness was 0.63 mm (n=4), 1 mm (n=14), 1.25 mm (n=28), 2.5 mm (n=75), or 5 mm (n=95). In-plane voxel dimensions ranged from 0.54–0.86 mm, and all reconstructed axial images had an in-plane matrix size of 512 by 512. The kVp setting for the scans was predominantly 120 kVp (n=212), with 100 kVp (n=1) and 140 kVp (n=3) also used. Reconstruction kernels fell into two large categories, with "Lung" kernels (including the Philips "L" and GE "Lung" kernels) used for 136 scans and "Standard" kernels (including Philips "B" and GE "chest," "soft," and "standard" kernels) used for the remaining 80 scans.

5.2.3 Lung Volume Quantification

Lung region segmentation was performed using the segmentation algorithm described by Sensakovic *et al.* [60] in the pleural disease segmentation scheme. The lung segmentation method is fully automated and utilizes gray-level, morphological, and texture features to segment the aerated lung tissue. The lung segmentation method has been under development in our group for some time, and the updated version was previously used in other studies for patients with MPM [107, 108]. The resulting segmentations were all reviewed for accuracy specifically for this study and modified when necessary by the author, who had been trained in thoracic anatomy by an attending physician. In-house software was used for this task (Abras), and duration of any necessary intervention was tracked.

A pixel-counting technique was used to calculate the volume of aerated lung tissue inside the resulting lung segmentations [94]. The number of pixels inside the lung segmentation were counted for each axial section and multiplied by the area of a single pixel and by the axial section thickness. These per-section lung volumes were summed for each axial section containing aerated lung; very occasionally, the axial section thickness changed over the span of a single scan, and therefore the volume of aerated lung needed to be summed on a section-by-section basis.

As an independent validation of the lung segmentation method, lung segmentations were performed on a separate set of 44 CT scans from 22 patients with MPM (one baseline and one follow-up scan per patient). For each patient, the full automated lung segmentation was performed, and an attending radiologist contoured the aerated lung tissue on three axial sections for the diseased (ipsilateral) lung and healthy (contralateral) lung (patients had unilateral disease). The area enclosed by each contouring method was calculated, and the section-by-section areas are compared using Pearson's correlation coefficient and Bland-Altman analysis [102].

Lung volumes were used as a response assessment measurement by normalizing the ipsilateral lung volume by the contralateral lung volume for each patient scan. While it is customary for CT scans to be acquired during patient breath-hold, it is possible that differences in patient respira-

tory phase between scan dates still exist. By using patients with unilateral disease, the healthy (contralateral) lung can be used to normalize the volume of aerated lung tissue in the diseased (ipsilateral) hemithorax, thereby controlling for any potential differences in inspired volume. This normalized volume V_{norm} is calculated as

$$V_{norm}(t_1) = \frac{V_{ipsilateral}(t_1)}{V_{contralateral}(t_1)}, \tag{5.1}$$

for a single timepoint t_1, and the specific growth rate (SGR) between baseline and t_1 is calculated from equation 4.5 as

$$SGR_{lung}(t_1) = \frac{\ln\left[\frac{V_{norm}(t_1)}{V_{norm}(t_0)}\right]}{(t_1 - t_0)}. \tag{5.2}$$

5.2.4 Data Analysis

The different tumor response assessment measurements in this study (summed linear thickness measurements, normalized lung volumes, and disease volumes) were compared using correlation statistics. The correlation between changes in disease volume and changes in normalized lung volumes was measured using Spearman's rank correlation, since while there is no geometric model for the physiological trade-off between lung volume and disease volume, any such model would be decidedly non-linear. To compare changes in summed linear thickness measurements with changes in disease volume, rank correlation was again used. While there do exist expressions relating changes in linear thickness to changes in volume for various geometric models (i.e., the mathematical models underlying the manuscript by Oxnard *et al.* [51]), only the sphere model could be tested in this study. The other geometric models (cylindrical annulus, crescent prism, and lens prism) require knowledge of the average linear thickness measurement and hemithorax diameter [51]; only the summed linear thickness measurement was available for this study (the number of linear thickness measurements varied clinically so the average value of a single thickness measurement was unattainable), and the hemithorax segmentations from the previous chapter were only drawn to

encapsulate the outer boundary of any present disease. While only obtainable for the sphere model (which might not be the most appropriate model for mesothelioma), the proportion of variability in observed volume changes explained by the geometry model is reported as R^2 from a linear fit between actual changes and predicted sphere model changes. For a sphere with diameter d that changes by an amount Δd, the volume V changes by

$$\frac{\Delta V}{V} = \left(\frac{\Delta d}{d}\right)^3 + 3\left(\frac{\Delta d}{d}\right)^2 + 3\left(\frac{\Delta d}{d}\right). \tag{5.3}$$

To compare the prognostic performance of the different disease response assessment measurements, the univariate significance of all three measurement techniques was assessed using time-varying Cox proportional hazards (PH) models, as in the previous chapter. Next, survival models were built using the clinical covariates from the final multivariate Cox PH model in Table 4.4 and one disease response assessment measurement, modeled as specific growth rate (SGR, see equation 4.5). Therefore, the clinical covariates disease histology, dyspnea, and Eastern Cooperative Oncology Group (ECOG) performance status were included along with (1) summed linear measurement SGR, (2) disease volume SGR, or (3) normalized lung volume SGR.

The performance of the survival models was assessed using the C^τ statistic introduced in the previous chapter. For this study, values of C^τ are reported from training and testing on the same dataset, as well as leave-one-out cross-validation (LOOCV) performance values for the different models. Additionally, repeated random sub-sampling was used to assess the difference between models. Since standard errors are not provided by the C^τ evaluation code in R, random sub-sample performance metrics can be used to assess the difference in paired patient cohorts for the models tested in this study. In each of 1000 sub-sample iterations, each model was trained on two-thirds of the patient cohort and tested on the remaining one-third of the patient cohort. The training set was chosen randomly without replacement at each iteration, and the testing set was considered to be the remaining patients who had not been selected for the training set at that iteration. Each model (full multivariate Cox PH model with summed linear thickness SGR, disease volume SGR, or

normalized lung volume SGR) was trained on the training cohort then tested on the testing cohort. Therefore, for each sub-sample iteration, model performance statistics are tracked in a *paired* fashion, and differences between models can be assessed using the histogram of paired differences between testing cohort performance values. Models were considered significantly different if the 95% central confidence interval (CI) of sub-sample paired differences did not include a difference of zero.

5.3 Results

5.3.1 Patients and Overall Survival

Median survival from baseline imaging prior to treatment initiation was 12.7 months (95% confidence interval, 10.2–15.3 months). Of the 61 patients, there were 58 observed deaths, while the remaining three patients were lost to follow-up after a median duration of 34 months. The overall Kaplan-Meier survival curve is shown in Figure 5.1.

5.3.2 Lung Segmentation

Across all patients, the mean baseline ipsilateral lung volume was 1021 ± 574 mL, and the mean baseline contralateral lung volume was 2648 ± 639 mL. The mean normalized ipsilateral lung volume at baseline was 0.399 (range 0.058–1.262). By the first follow-up scan, the normalized ipsilateral lung volume had increased to 0.420, up by a geometric mean of 5% from baseline. By the end of treatment, the normalized ipsilateral lung volume had increased an average of 8% from baseline. Over the course of the entire treatment, the distinction between normalized ipsilateral lung volume increase and decrease was significantly associated with patient survival. Figure 5.2 shows the Kaplan-Meier survival curves for the two patient groups, and the log-rank test for differences in survival showed the groups to be significantly different ($p = 0.0003$).

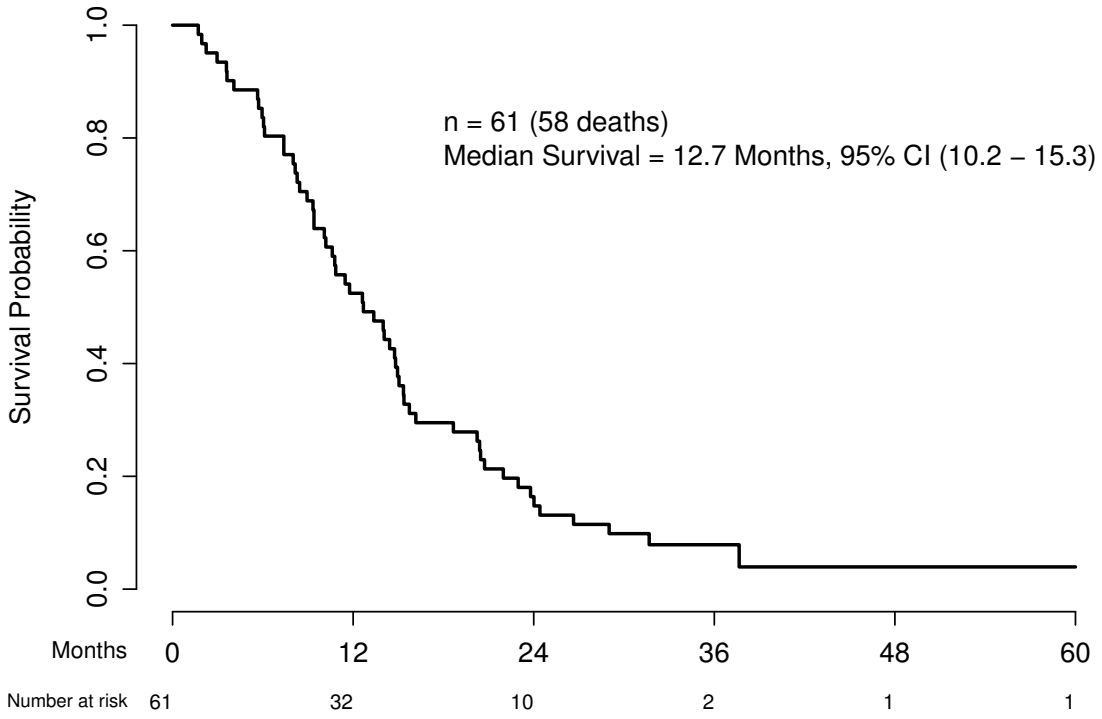


Figure 5.1: Overall survival for the patient cohort in this study.

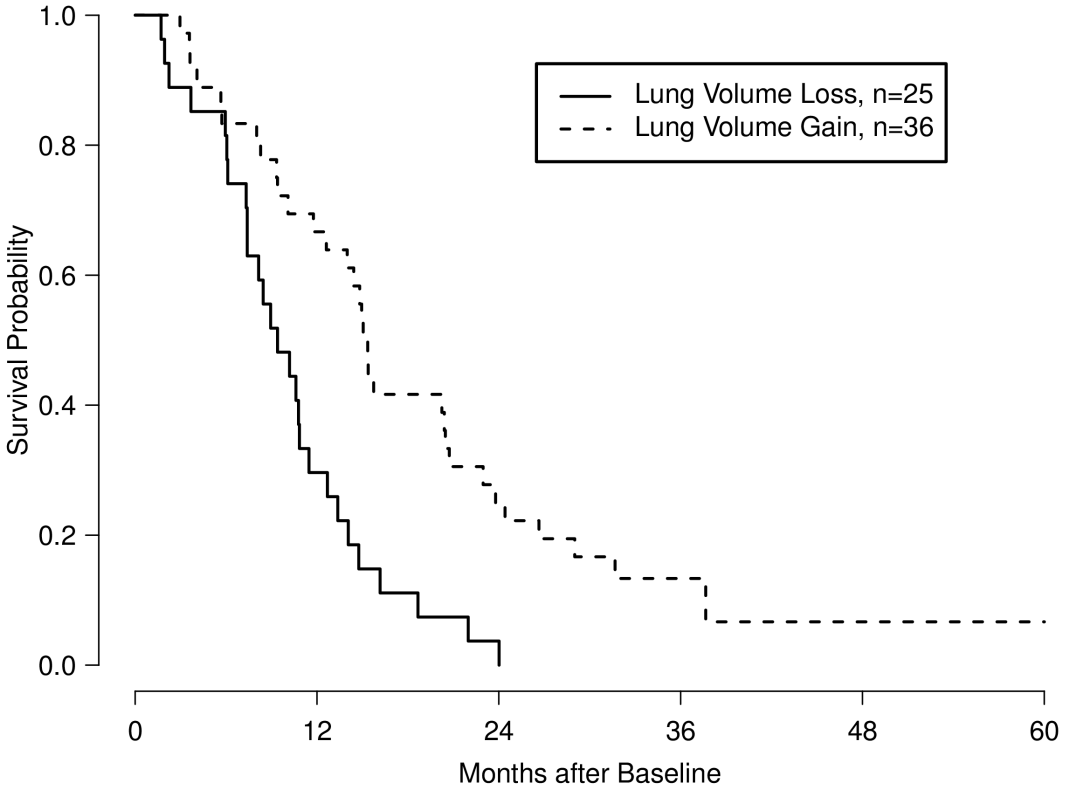


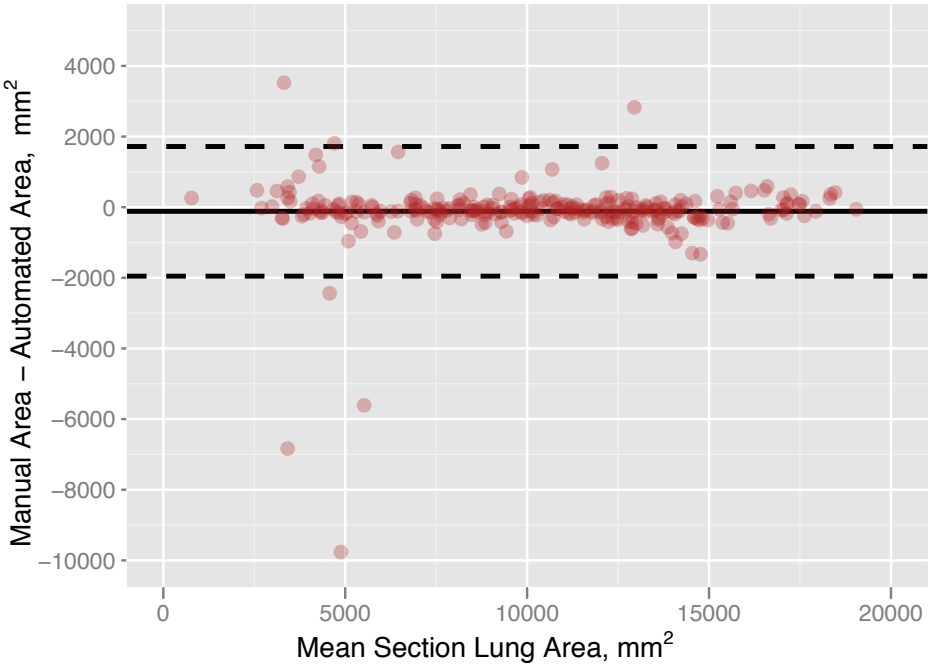
Figure 5.2: Kaplan-Meier survival curves for patients with and without normalized ipsilateral lung volume increase during the course of their therapy.

The extent of manual intervention necessary in the otherwise fully automated lung segmentation was minimal. For cases that required any intervention whatsoever (21% of all scans), the duration of manual intervention averaged approximately one minute. Only 1.9% of cases required five minutes or more of manual intervention. The predominant cause for manual editing of lung segmentations was erroneous inclusion of segmented bowel gas.

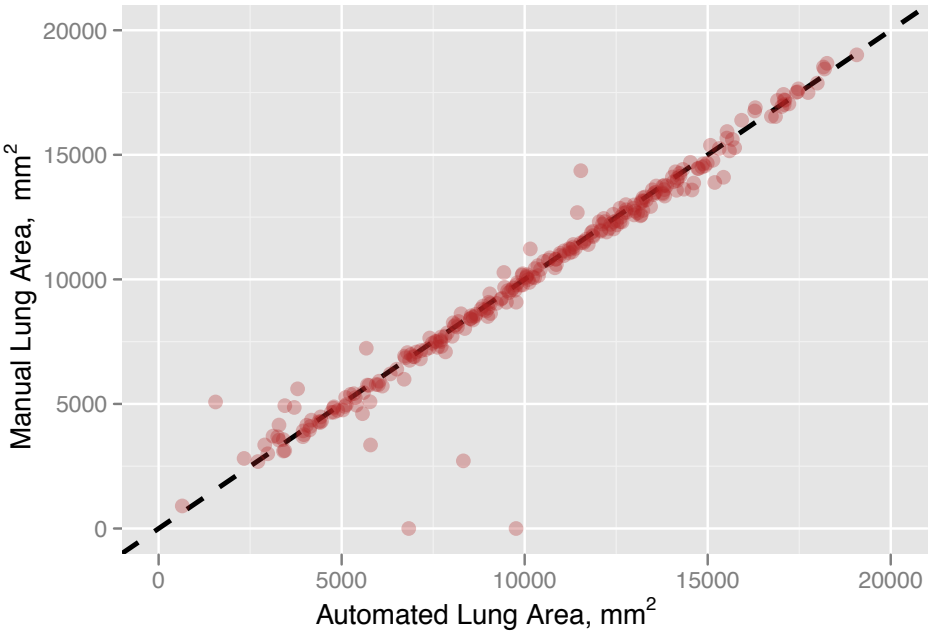
From the validation study, there was very high agreement for area measurements of per-section lung segmentations between the manual approach and the automated method for the 132 axial sections evaluated. Pearson's correlation coefficient was calculated as 0.973 ($p < 0.0001$). However, because the intent of both measurement techniques was the quantification of the *same structure*, one would expect a very high correlation (in fact, only a small correlation coefficient would be noteworthy). Using Bland-Altman analysis, the mean bias was for automated measurements to be 117 mm^2 larger than manual measurements (the average section lung area was 10203 mm^2), or 1.1% larger on average. The 95% limits of agreement in the difference between manual measurements and automated measurements were -1952–1719 mm^2, relatively small given the correlation and average measurement magnitude. The Bland-Altman plot is shown in Figure 5.3.

5.3.3 Linear and Volumetric Measurement Correlations

A plot comparing the relative change from baseline of summed linear thickness measurements and disease volumes computed according to the semi-automated methods in Chapter 4 for the 61 patients in this study is shown in Figure 5.4. Each of the 155 points on the plot represents a single paired change from baseline (i.e., if a patient has four CT scans, there will be three data points comparing linear measurements with volume measurements for that patient in Figure 5.4). For these data, Spearman's rank correlation coefficient was estimated as $\rho_{thickness} = 0.676$ ($p < 0.0001$) and Pearson's linear correlation coefficient was estimated as $r_{thickness} = 0.665$ ($p < 0.0001$). Both correlations are positive, as expected, indicating that growth in disease linear thickness corresponds to growth in disease volume.



(a)



(b)

Figure 5.3: Validation of automated lung segmentation. 5.3a, Bland-Altman plot, where bias is shown with a solid black line and the 95% limits of agreement are shown with dashed black lines. 5.3b, direct comparison between measurements, with the identity line shown.

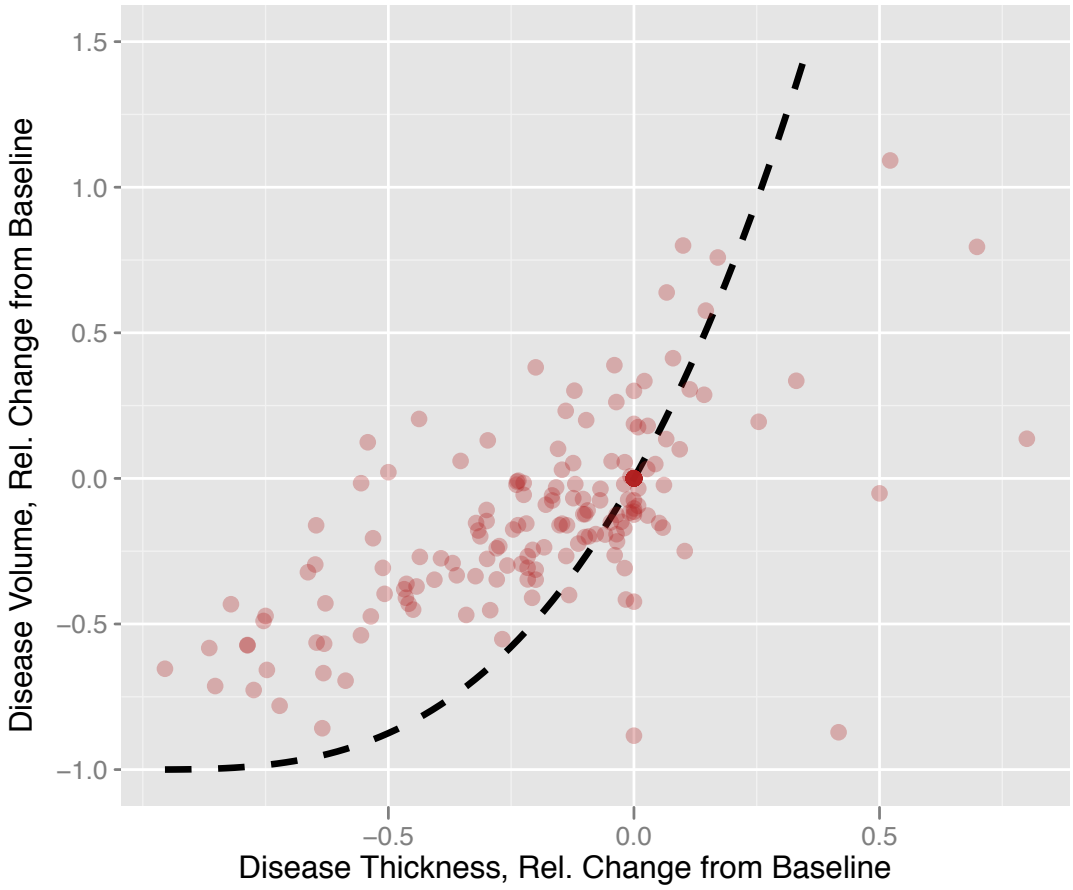


Figure 5.4: Relative change from baseline of summed linear thickness measurements versus relative change from baseline of disease volumes. The relationship expected from a spherical geometric model is indicated with a dashed black line.

The relationship expected from a spherical geometric model is indicated in the plot with a dashed line (the relationship is given by equation 5.3). The quality of fit of the spherical model is $R^2 = 0.35$, indicating that only 35% of the overall variation in the data is attributable to the geometric model. Visual inspection of the plot indeed indicates that the data do not reliably fall along the dashed line and instead appear nearly linear in some locations. While there was no theoretical reason to believe that mesothelioma would follow a spherical geometry, Figure 5.4 provides the first empirical evidence for the inappropriateness of the spherical assumption implicit in the standard RECIST classification criteria for MPM.

A plot comparing the relative change from baseline of normalized ipsilateral lung volumes and

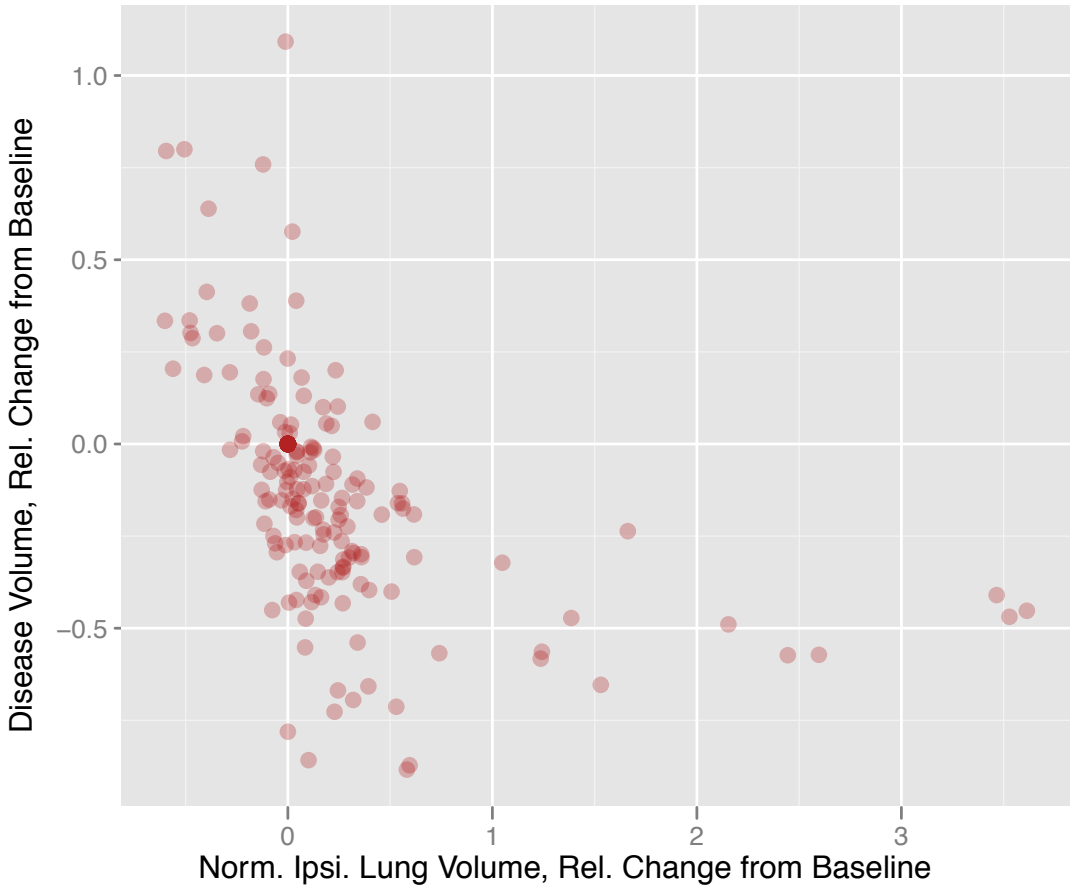


Figure 5.5: Relative change from baseline of normalized ipsilateral lung volumes versus relative change from baseline of disease volumes.

actual disease volumes computed according to the semi-automated methods in Chapter 4 for the 61 patients in this study is shown in Figure 5.5. Each of the 155 points on the plot represents a single paired change from baseline (i.e., if a patient has four CT scans, there will be three data points comparing normalized ipsilateral lung volumes with disease volumes for that patient in Figure 5.5). Using the non-parametric Spearman's rank correlation coefficient, the correlation was estimated to be $\rho_{lung} = -0.687$ (testing for difference from zero, $p < 0.0001$). The linear trend correlation from Pearson's correlation coefficient was estimated to be $r_{lung} = -0.494$ ($p < 0.0001$). As expected, the correlation coefficients are negative, indicating that for an increase in normalized lung volume, the disease volume decreases.

5.3.4 Survival Analysis

All three response assessment measurement techniques were significantly associated with patient survival in univariate Cox PH survival models. Increases in continuous time-varying linear thickness SGR measurements were associated with poor patient prognosis (HR=1.53, $p < 0.0001$), as were increases in disease volume SGR (HR=1.32, $p = 0.0003$). As expected, decreases in normalized ipsilateral lung volume SGR were associated with poor patient prognosis (HR=0.763, $p = 0.003$).

In multivariate Cox PH survival models including disease histology, dyspnea, and ECOG performance status, all three response assessment measurement techniques remained significantly associated with patient survival. The model coefficients for the summed linear thickness model, disease volume model, and normalized lung volume model are shown in Tables 5.2, 5.3, and 5.4, respectively. Note that the clinical covariates included in these models were selected from the final multivariate model in the previous chapter and were not necessarily selected for their prognostic significance in this particular patient cohort.

The hazard ratio estimates for the clinical covariates vary between the three tumor measurement technique models, but the variability is small compared to the 95% confidence intervals given in Tables 5.2, 5.3, and 5.4. For instance, the hazard ratio for "Other" histology is 1.95 times higher than for Epithelioid histology when tumor burden is modeled using summed linear thickness measurements, but the same comparison shows that "Other" histology is 2.38 times worse than Epithelioid histology for patient mortality when tumor burden is modeled using normalized ipsilateral lung volume measurements. It is possible that the effect of disease histology is simply more pronounced in the lung volume model, but given the overlap in the 95% confidence intervals of the hazard ratio estimates one cannot state that the effect of histology is *significantly* different between the two models. The same can be said for the other clinical covariates as well. It is interesting to note that in all three multivariate models the tumor response measurement technique is always the most significant of the prognostic variables. While the values of the hazard ratios are different

Variable		Hazard Ratio	95% CI	p-value
Summed Linear Thickness	(continuous, SGR)	1.468	[1.18, 1.84]	0.00053
Histology	Epithelioid	1	–	–
	Other	1.953	[1.05, 3.65]	0.036
Dyspnea	No	1	–	–
	Yes	2.458	[1.09, 5.55]	0.030
ECOG Performance Status	0	1	–	–
	1 or 2	1.472	[0.83, 2.61]	0.099

Table 5.2: Multivariate Cox PH model, including hazard ratios and 95% confidence intervals (CI). Summed linear thickness modeled as continuous specific growth rate (SGR) from baseline.

for the three measurement techniques (as they should be, given the differences in the structures being measured), the p-value for the tumor response measurement technique is always smallest, indicating in some sense that the tumor response measurement technique is the most "important" prognostic variable for all three models.

Model performance was quantified using the C^{τ} statistic, where τ was set to the time associated with the final observed patient death (37.7 months). The performance of the full multivariate model trained and tested on the same patient cohort was 0.692, 0.680, and 0.670 for the models using summed linear thickness measurements, disease volume measurements, and normalized lung volume measurements, respectively. In the leave-one-out cross-validation, these scores were reduced somewhat, as expected, to 0.657, 0.625, and 0.630, respectively. Finally, the mean random sub-sample performance values for the three models were 0.659, 0.638, and 0.628, respectively. These values are summarized in Table 5.5.

To compare one response assessment measurement technique with another, the paired differences in sub-sample testing cohort performance values are reported. The mean difference in paired C^{τ} performance values between the summed linear thickness model and the disease volume model

Variable		Hazard Ratio	95% CI	p-value
Disease Volume	(continuous, SGR)	1.334	[1.13, 1.58]	0.00090
Histology	Epithelioid	1	–	–
	Other	2.041	[1.10, 3.79]	0.023
Dyspnea	No	1	–	–
	Yes	2.809	[1.19, 6.61]	0.018
ECOG Performance Status	0	1	–	–
	1 or 2	1.543	[0.89, 2.67]	0.12

Table 5.3: Multivariate Cox PH model, including hazard ratios and 95% confidence intervals (CI). Disease volume modeled as continuous specific growth rate (SGR) from baseline.

Variable		Hazard Ratio	95% CI	p-value
Normalized Lung Volume	(continuous, SGR)	0.762	[0.64, 0.91]	0.0033
Histology	Epithelioid	1	–	–
	Other	2.375	[1.30, 4.34]	0.0050
Dyspnea	No	1	–	–
	Yes	2.154	[0.98, 4.74]	0.056
ECOG Performance Status	0	1	–	–
	1 or 2	1.582	[0.92, 2.73]	0.099

Table 5.4: Multivariate Cox PH model, including hazard ratios and 95% confidence intervals (CI). Normalized lung volume modeled as continuous specific growth rate (SGR) from baseline.

	"Full" Performance	LOOCV Performance	Mean Random Sub-Sample Performance	95% Random Sub-Sample Confidence Interval
Summed Linear Thickness	0.692	0.657	0.659	[0.556, 0.760]
Disease Volume	0.680	0.625	0.638	[0.526, 0.755]
Normalized Lung Volume	0.670	0.630	0.628	[0.510, 0.744]

Table 5.5: Performance value (C^{τ}) summary for multivariate survival models from Tables 5.2, 5.3, and 5.4. Performance values are given for "full" models trained and tested on the complete cohort, from leave-one-out cross-validations, and from repeated random sub-sample simulations.

was 0.022, with a 95% confidence interval of -0.077–0.123 and was therefore not significant (bootstrap $p = 0.30$). The mean difference in paired C^{τ} performance values between the normalized ipsilateral lung volume model and the disease volume model was -0.009, with a 95% confidence interval of -0.087–0.077, and was therefore not significant (bootstrap $p = 0.65$). Figures 5.6 and 5.7 show histograms of the paired C^{τ} differences in model performance values for summed linear thickness versus disease volume and for normalized ipsilateral lung volume versus disease volume, respectively. While the performance of the summed linear thickness model is on average 3.4% higher than the performance of the disease volume model, there is obviously considerable overlap in the performance of the two models. The disease volume model outperformed the linear thickness model for 30% of the random sub-sample iterations. For the normalized ipsilateral lung volume model, though, the performance is on average 1.4% lower than the performance of the disease volume model. There is even more overlap between the lung and disease volume models than the linear thickness and disease volume models. For the purposes of survival prediction, changes in normalized ipsilateral lung volume are statistically very similar to changes in disease volume,

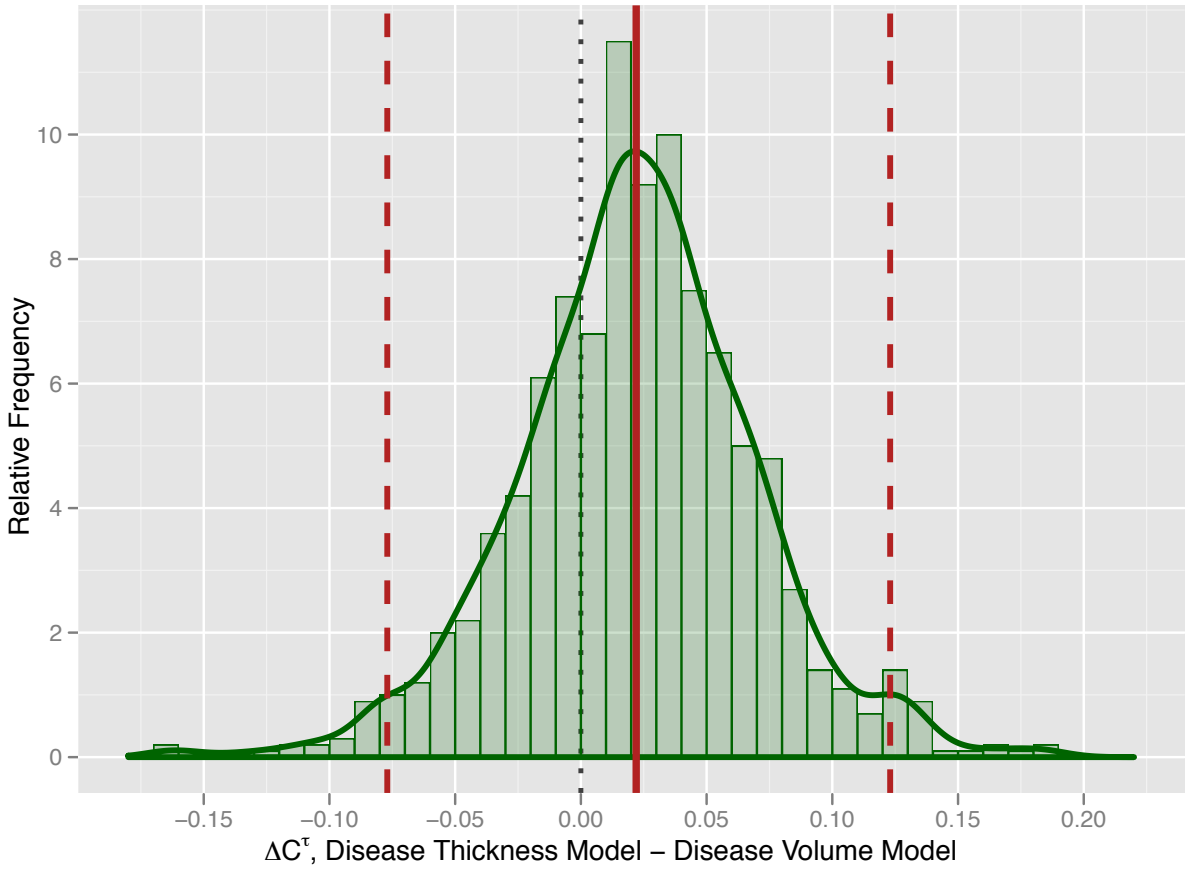


Figure 5.6: Smoothed histogram of repeated random sub-sample performance differences, ΔC^{τ}, comparing multivariate Cox PH models using summed linear thickness measurements and disease volume measurements. All models were trained on two-thirds of the patient cohort selected at random and tested on the remaining third (1000 iterations). Mean and 95% confidence interval of differences shown with red lines.

as indicated in Figure 5.7.

5.4 Discussion

In the previous chapter, it was shown for the first time that continuous and time-varying image-based measurements of pleural disease volume were significantly associated with patient survival. The performance of the survival model was even improved with the addition of other clinical co-variates to the model. This chapter tells a largely similar story, though now the protagonist role is

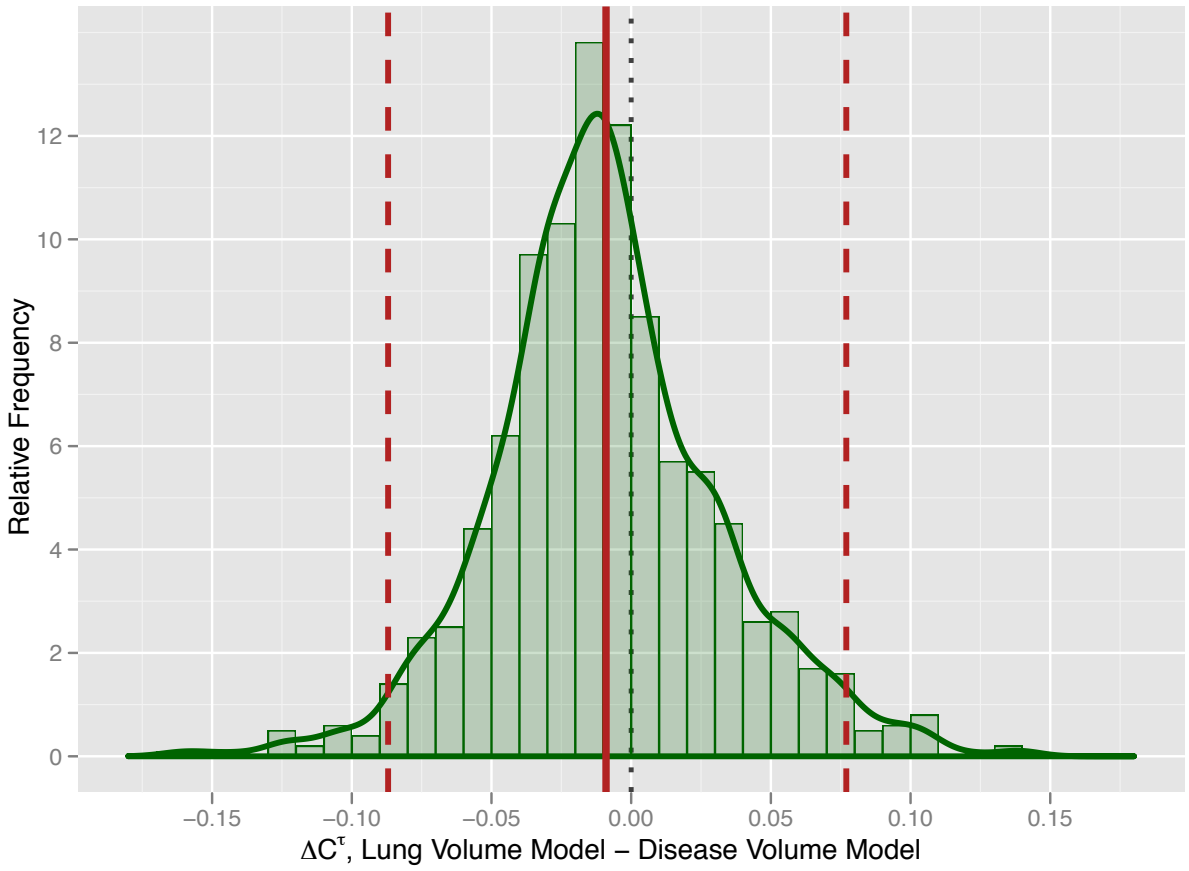


Figure 5.7: Smoothed histogram of repeated random sub-sample performance differences, ΔC^{τ}, comparing multivariate Cox PH models using normalized ipsilateral lung volume measurements and disease volume measurements. All models were trained on two-thirds of the patient cohort selected at random and tested on the remaining third (1000 iterations). Mean and 95% confidence interval of differences shown with red lines.

shared between three tumor response assessment techniques: summed linear thickness measurements acquired using the modified RECIST technique, semi-automated segmentations of pleural disease volume, and automated segmentations of normalized ipsilateral lung volume. These three response assessment measurement techniques are all significantly associated with patient survival, and there are no significant differences between model performance metrics. However, there are certainly practical differences between the three measurement techniques and the models derived therefrom.

In a way, measurements of disease volume are the most established for use in response assessment. After all, response assessment as performed by RECIST is historically based on equivalent *volumetric* response from simplified measurements. Until recently, though, full measurements of pleural disease volume for patients with MPM were time-prohibitive, and summed linear thickness measurements remain the clinical standard for response assessment. In the past few years, several software algorithms for the segmentation of mesothelioma on CT scans have been published [57, 58, 60], and researchers are now able to explore true disease volume as a response assessment measurement technique. The novel response assessment technique in this study is lung volume: aerated lung segmentation is a comparatively easier computational task than pleural disease segmentation, and there is reason to expect lung volumes to be uniquely correlated physiologically to disease volumes for patients with MPM. Because of the constrained hemithoracic space, any decrease in disease volume should be met with corresponding increase in the ipsilateral (disease-sided) lung volume. Normalizing the ipsilateral lung volume by the contralateral lung volume corrects for differences in respiratory phase between CT scans, and the resulting normalized lung volume forms a useful (if contrary) response assessment technique.

The correlations among the three measurement techniques are in line with expectations. One would expect changes in summed linear thickness to be correlated with changes in disease volume, and the data show this to be true. However, the spherical geometric relationship implicit in the RECIST protocol is not a good fit, as evidenced in Figure 5.4. Unfortunately, the alternate

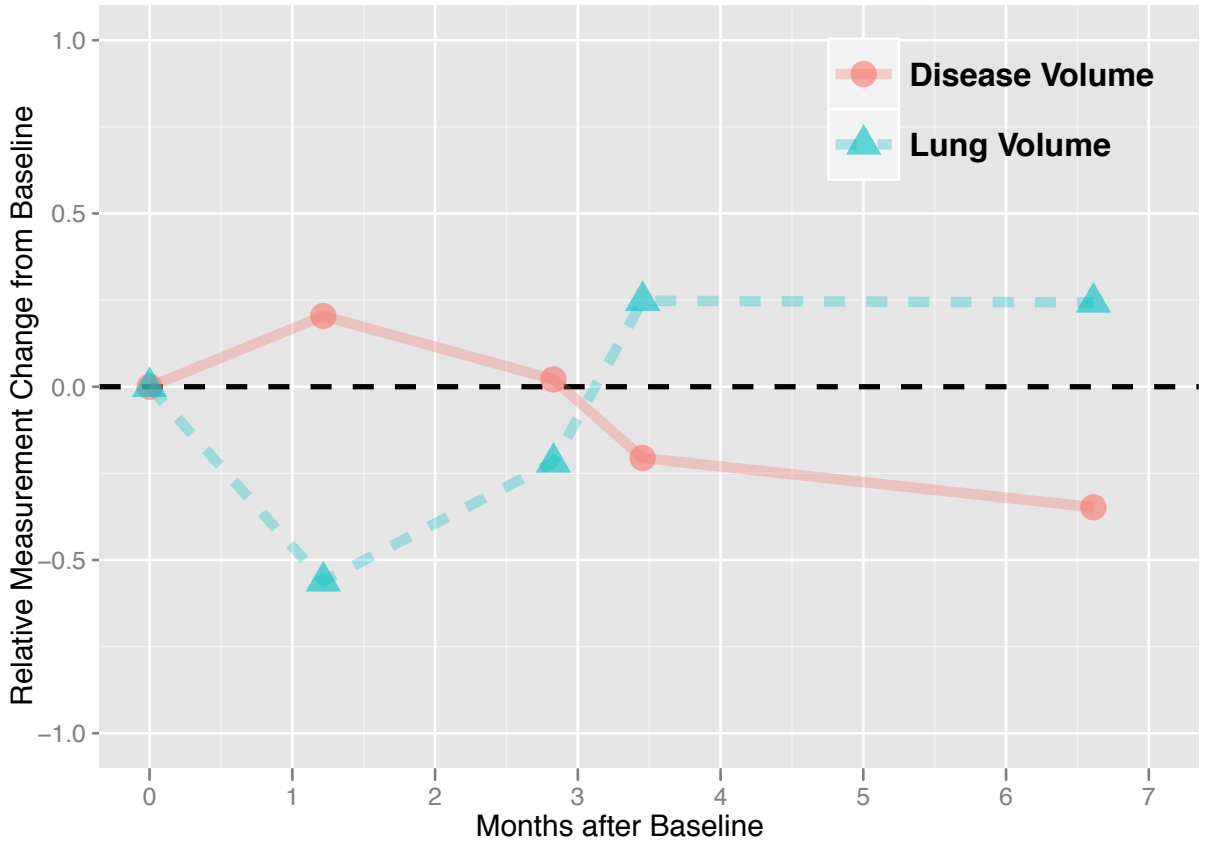


Figure 5.8: Relative changes from baseline of normalized lung volume and pleural disease volume for an example patient. Note the (anti-)correlation between the two curves.

geometric models from Oxnard *et al.* [51] could not be tested in this study because of the form of available data. The correlation between normalized ipsilateral lung volumes and disease volumes was also as expected, since increases in normalized ipsilateral lung volume were met by decreases in disease volume. An example of this correlation is shown in Figure 5.8, where changes in normalized ipsilateral lung volume and changes in disease volume are seen to closely mirror one another.

It is possible to build multivariate survival models simultaneously utilizing any two (or even all three) tumor response measurement techniques. However, because of the high correlation between the three tumor response measurement techniques, using more than one measurement technique in the same multivariate Cox PH model results in at least one of the tumor response measurement

terms becoming non-significant (usually with a p-value larger than 0.20). Therefore, no more than one tumor response measurement technique at a time can be an independent significant covariate for patient prognosis. This "only one at a time" observation further supports the idea that all three measurement techniques are inherently measuring the same underlying process: changing tumor burden.

The performance of the survival model using pleural disease volumes in this study was near the performance reported in the previous chapter. The slightly lower performance may be attributable, in part, to the reduction in patient cohort size for this study, especially since the size of each random sub-sample testing cohort fell from 27 to 20 patients. The more interesting comparison is between survival models using alternate disease response measurement techniques. The fact that the survival model with summed linear thickness measurements outperformed the disease volume survival model was unexpected. While the survival model using manual thickness measurements did not perform significantly better than the disease volume model, the performance was better on average, as evidenced by Figure 5.6. Disease volumes are logically better able to capture changes in overall tumor bulk, but perhaps changes in tumor thickness are physiologically more predictive of eventual patient survival than overall volumetric changes. The two measurement techniques provide different information, and while it was previously assumed that disease volumes should be the ultimate goal of any response assessment technique, it is possible that the specific type of morphological change quantified by tumor thickness measurements is simply more representative of patient benefit. Another (purely speculative) possibility is that human observers are able to place their baseline tumor thickness measurements in locations that are in some sense "important" for response assessment; volume measurements capture changes over the total extent of disease, while summed tumor thickness measurements only capture change in the discrete (up to six, by modified RECIST) locations where baseline measurements were placed.

Also interesting is the nearly statistically identical performance of the survival models using disease volumes and normalized ipsilateral lung volumes (see Figure 5.7). The similar performance

of the two models reinforces the expectation that changes in (normalized) lung volume and disease volume should convey roughly equivalent information due to the physiological correlation between the two structures. The correlation between paired C^{τ} values from random sub-sample testing cohorts showed high correlation ($r = 0.77$) between the survival models using disease volume and normalized ipsilateral lung volume, again as expected.

There are various advantages and disadvantages for each response assessment measurement technique. It was shown by Frauenfelder *et al*. [57] that the inter-observer variability is substantially lower for disease volume measurements than for linear thickness measurements, a fact that could become an important consideration if disease volumes were to be used clinically to assess tumor response. However, linear measurements require less manual intervention time than disease volume measurements, and existing techniques could potentially be used to partially automate the linear measurement process and thereby reduce variability [52, 53]. Lung volume measurement is an automated process, and the only manual intervention used in this study was the correction of obvious segmentation errors from contrast artifacts and bowel gas. It is therefore reasonable to believe that lung volume measurements would have almost no inter-observer variability whatsoever. However, the utility of lung volume measurements for tumor response assessment is limited to patients with unilateral disease. While unilateral disease is common (the number of useable patients in this study would have increased only from 61 to 68 with the removal of the "unilateral disease" criterion), this stipulation necessarily precludes response assessment via lung volumes for a small number of patients.

An inherent limitation of this study is the relatively small number of patients used. The survival models compared in this study form a starting point for a validation in independent patient cohorts and should not be taken as "definitive" response models. While all the survival models in this study are significant (all of the 95% random sub-sample confidence intervals for C^{τ} do not include 0.5), absolute performance scores of around 0.65 are by no means perfect. And while the survival model from the linear thickness measurements outperforms the other two models on average, there is no

statistical basis to conclude that any one model is better than another. It should be further cautioned that the survival models in this study may not be applicable to patients who receive biologically different treatments than the cytotoxic therapy used herein.

In summary, survival models using three different tumor response measurement techniques were compared in this study for patients with MPM undergoing chemotherapeutic treatment. Models were fit using clinical covariates identified in a previous chapter and either summed linear thickness measurements, pleural disease volume measurements, or normalized ipsilateral lung volume measurements. As a novel tumor response assessment technique, lung volumes exhibited the expected correlation with disease volumes. All three tumor response measurement techniques were significantly associated with patient survival. The model using summed linear thickness measurements performed, on average, better than the other two models, though the differences were not significant.

CHAPTER 6

DYNAMIC COMPUTED TOMOGRAPHY FOR MESOTHELIOMA

"X rays will prove to be a hoax." – Sir William Thomson (Lord Kelvin)

6.1 Introduction

Standard computed tomography (CT) captures a volumetric image of an object over the course of a relatively small amount of time. These single-timepoint images contain structural anatomical information, where the gray-scale value associated with each discrete reconstructed volumetric pixel (voxel) is related to the mean x-ray attenuation coefficient of the materials in that voxel. These Hounsfield Unit (HU) values are assigned in relation to the individual voxel attenuation coefficients and the known attenuation of water. The addition of iodinated contrast media aids in the visualization of structures that take up contrast through the blood supply, since the x-ray attenuation of iodine is considerably higher than usual soft tissues for the x-ray energies under consideration (on the order of 100 keV). Therefore, regions that have contrast uptake will appear "brighter" on reconstructed images and will thus convey some physiologic information along with anatomic information. Depending on the time delay between contrast injection and volumetric image acquisition, the relative enhancement of different anatomical structures will imply some information about the hemodynamic properties of those structures. Contrast is used clinically in CT to both visually enhance structures (especially mediastinal structures) and identify physiological properties of interest.

Even with the addition of iodinated contrast, CT images pose challenges for the anatomic identification of disease for patients with malignant pleural mesothelioma (MPM). For reasons identified previously, MPM can be difficult to impossible to distinguish from surrounding soft tissues, and this fact has been quantitatively established in a previous study [24]. But a traditional contrast-enhanced scan is like a photograph, just a snapshot in time: an amount of contrast appropriate for

the patient (usually 90–120 mL of contrast with an iodine concentration of 350 mg/mL) is injected intravenously, and the CT scan is acquired after an amount of time that allows for distribution of the contrast media throughout the body (usually 30–90 seconds). On the other hand, much more physiological or functional information can be gathered from dynamic contrast-enhanced (DCE) imaging, which is more like a movie than a photograph. The main difference with DCE imaging is that instead of acquiring a single timepoint scan after a set delay from initiation of injection, scans are repeatedly captured during (and potentially following) the intravenous injection of the contrast media. In this way, DCE imaging adds a temporal component to an otherwise static structural scan. Because the change in HU value for a given voxel is directly proportional to the contrast of iodine in that voxel, blood flow to different tissues can be tracked over time from the temporal scan information. If a single timepoint contrast-enhanced scan gives information about relative vascularity between various soft tissues, a DCE scan gives information about how quickly the blood is flowing to the tissues of interest, how much blood is flowing through those tissues, and how long the blood takes to travel through those tissues. For clarity, while DCE-CT is often called "perfusion CT," the author has purposefully restricted the use of the term "perfusion" to mean a specific hemodynamic parameter calculated from dynamic imaging data.

The individual voxel contrast uptake curves can be analyzed to extract various summary hemodynamic properties. The hemodynamic maps of blood flow and vascularity have been shown to serve as useful oncologic biomarkers [75]. In a study using DCE-MRI to image MPM patients, Giesel *et al.* [76] reported that a hemodynamic perfusion parameter (k_{ep}) was clearly associated with patient survival. However, the general application of DCE-MRI for MPM is limited by the presence of cardiac and respiratory motion. It is therefore reasonable to believe that DCE-CT will produce functional data maps that hold prognostic significance in a manner similar to DCE-MRI imaging. And while DCE-CT uses ionizing radiation to generate scan data, the scans do not suffer from the MRI artifacts mentioned previously, and the DCE-CT scan can be integrated efficiently with the current clinically indicated CT scanning already being performed on MPM patients.

To our knowledge, DCE-CT has been used for imaging of MPM in only a single study conducted by Meijerink *et al*. [80]. The study investigated the use of DCE-CT to assess the hemodynamic response of tumors of the thorax and abdomen. The patient cohort consisted of 16 patients, including only two patients with MPM. DCE-CT data demonstrated a substantial decrease in tumor perfusion values for both MPM patients after treatment with anti-angiogenic VEGF inhibitors. This supported the assertion that DCE-CT is able to measure hemodynamic function in MPM patients.

In order to more completely explore the potential use of DCE-CT for patients with MPM, we initiated a pilot study to explore the feasibility of DCE-CT and the information contained in the hemodynamic maps resulting from the DCE-CT scans. Besides establishing feasibility for DCE-CT performed in a slightly different manner than usual, the study also aimed to identify the "optimal" delay period between contrast injection and peak enhancement of the tissues of interest. Finally, the pilot study was structured to obtain scans at two distinct times for each patient, thereby leading to an analysis of changes in hemodynamic data maps over time and correlations between changes in DCE-CT data and tumor bulk. This pilot study is still underway (May 2012), with an accrual goal of 20 patients, each scanned with DCE-CT twice (40 scans total). Ten patients will be "on observation," not receiving any biologically active therapy, and the remaining ten patients will be "on treatment," though the (chemotherapeutic) treatment is not necessarily the same among all patients. There are, at present, 13 patients with two scans each, and another three patients who have received only their first scan. Of the 13 patients with two scans each, seven are on observation, and all three patients awaiting their second scan are on observation as well. The report that follows is a documentation of the imaging protocol used, the image processing and other methodologies required for the analysis of the DCE-CT data, and the quantitative results obtained thus far for the 13 patients with two scans each.

6.2 Imaging Protocol

6.2.1 Patient Cohort

Patients were considered eligible for this study approved by our local Institutional Review Board if they had been diagnosed with MPM with pathologic confirmation. Patients were required to have a tumor with a thickness of at least 1 cm located axially between the superior-most point of the aortic arch and the superior-most point of the diaphragm on a previous chest CT examination, a criterion ensuring that patients would have a volume of disease sufficient for quantitative analysis. Patients were only eligible if they were not considered potential candidates for surgical intervention during their involvement on the study and could not have undergone talc pleurodesis in the three months prior to the first DCE-CT scan. Surgical intervention during the study would obviously remove foci of disease, leading to an inability to compare uptake curves and changes in hemodynamic properties between the two scan dates. Talc pleurodesis is known to be associated with inflammation and other local immunological changes and would potentially alter the hemodynamic properties of tissue. In such situations, any changes in DCE-CT data between the two scan dates could be an arbitrary mixture of effects from talc pleurodesis healing and effects from any active therapy. As mentioned previously, the overall patient accrual goal was 20 patients, with 10 patients each from the observation and treatment groups. Finally, and perhaps most importantly, patients were only eligible for this study if they were scheduled to receive their clinically indicated standard chest CT scan with contrast. Some patients do not tolerate contrast media, depending on kidney function, and therefore our study only considered patients who would have otherwise been given the full amount of clinically indicated contrast. All patients provided written informed consent for their participation in this prospective study.

For the six patients in this study "on treatment," the treatment regimens were not identical across patients. Two patients were on a clinical trial for oral Vorinostat, a potential new inhibitor for malignant pleural mesothelioma [122]. Two patients were on a clinically standard regimen

of cisplatin and pemetrexed. One patient received cisplatin, pemetrexed, and CBP501, a novel treatment which interferes with the cellular reproduction cycle at the G_2 checkpoint [123]. Finally, one patient received pemetrexed only. From the patients on observation, one patient was part of the placebo arm of the Vorinostat trial. These details are provided in Table 6.1.

The time between patient scans was usually approximately six weeks for patients undergoing treatment, since this is the time span of two cycles of chemotherapy. Treatment patients are typically scanned with CT every other cycle of chemotherapy, and the timing of the two scans in this study was set according to the clinical need for the standard chest CT component. The median duration between scans was 42 days for treated patients (range, 42–56 days). For patients on observation, the timing of the two scans was again mandated by the clinical need for the standard chest CT component. The median duration between scans was 58 days for patients on observation (range, 41–126 days). At the date of this analysis, four patients have died, with survival durations from diagnosis to death ranging from 9.2–71.1 months. The patients who remain living have a median follow-up duration from diagnosis of 31.9 months (range, 16.5–61.7 months).

6.2.2 DCE-CT Scan Acquisition

A DCE-CT scan starts the same way that any contrast-enhanced CT begins: a bolus of iodinated contrast media is injected venously via a power injector. The contrast agent enters the circulatory system and perfuses to different tissue types according to their vascularity and blood flow, where the increased attenuation of iodine will present with increasing HU value [77]. The degree of tissue enhancement depends on the contrast media concentration and local hemodynamics reflecting tissue vascularity and physiology [78]. During the DCE-CT scan, a small section of relevant patient anatomy (limited axially by the extent of the collimated CT beam) is repeatedly imaged, and the uptake and eventual washout characteristics of the iodinated contrast can be quantified by tracking voxel-by-voxel HU values over time.

The imaging protocol used specifically in this study had some constraints not typically present

Patient	Sex	Age	Treatment Status
1	Male	81	Observation (Placebo)
2	Male	52	Treatment (Vorinostat)
4	Male	80	Observation
5	Male	59	Observation
7	Female	64	Treatment (Cisplatin, Pemetrexed)
8	Male	77	Observation
10	Male	84	Treatment (Vorinostat)
12	Male	77	Treatment (Cisplatin, Pemetrexed)
13	Male	64	Treatment (Cisplatin, Pemetrexed, CBP501)
14	Male	79	Observation
15	Male	77	Observation
17	Male	68	Observation
21	Male	68	Treatment (Pemetrexed)

Table 6.1: Summary information about the 13 patients included in this study to date.

in usual DCE-CT imaging, since this protocol needed to fit in the workflow of an already ordered contrast-enhanced chest CT scan. The initial funding for this study was provided by the Hodges Society of the Department of Radiology at The University of Chicago and was not sufficient to provide contrast media and scanner time if the scan was not covered by insurance. Furthermore, it is desirable to minimize the number of iodinated contrast injections for a given patient due to the possibility of renal damage. Therefore, an imaging protocol was needed that would somehow utilize the bolus of contrast already necessary for the standard contrast-enhanced chest CT scan and would minimize extra scanner time. Furthermore, in clinical practice the standard chest CT scan usually occurs somewhere around 60–70 seconds post contrast injection initiation, so again this was a requirement during protocol planning.

The scanner used in this study is a Philips Brilliance iCT 256-slice scanner. After adjusting the reconstructed field-of-view to encompass an average patient's thorax, the maximum axial extent of the CT detector at isocenter is approximately 55 mm. Since individual DCE-CT acquisitions, referred to herein as "snapshots," are typically acquired without the CT table moving during the acquisition, we are limited in the amount of disease we can image for a given scan. While the Philips scanner is capable of table motion during snapshot acquisition, the table top does not move quickly compared with the CT gantry rotation time (0.3 seconds per rotation), and therefore extending the axial extent of snapshot imaging would generally lead to unacceptable motion (cardiac and respiratory). Before each DCE-CT scan, then, an anatomical location was preselected for DCE-CT imaging based on the patient's previous CT examinations (the location to scan was selected for all patients by the author). When selecting this 55 mm DCE-CT section, preference was given to areas of thicker disease and more superior axial locations to minimize respiratory motion. The DCE-CT section was constrained to be between the top of the aortic arch and the top of the diaphragm. The DCE-CT section for a patient's second DCE-CT scan was matched as closely as possible to the first scan.

The first component of our imaging protocol is an anteroposterior scout scan of the thorax

and abdomen. This is a standard component of any chest CT scan and is used to identify the location to scan for the rest of the CT series. The DCE-CT scan components are planned based on the anatomical location of the DCE-CT section identified previously, usually with reference to the radiologic position of the bifurcation of the trachea. Next, the contrast media is prepared and loaded into a power injector. Injection is typically performed via an 18-gauge cannula placed in the antecubital vein. Finally, the CT table is moved to the axial position indicated for full coverage of the selected DCE-CT section.

The timing of the protocol designed for this study is summarized in Figure 6.1. Contrast injection is started at time $t = 0$, concurrent with the first DCE-CT snapshot. Snapshots are acquired every three seconds until 20 snapshots have been acquired after 57 seconds. During this period, patients are instructed to use "shallow breathing," where they breathe enough to maintain comfort but avoid any sudden large inhalations or exhalations. After the 20th DCE-CT snapshot is acquired, the scanner transitions to the acquisition of the full clinical chest CT scan. This process usually requires between 10 and 18 seconds, depending on the axial location of the DCE-CT scanning section. Therefore, the full chest CT usually begins around 75 seconds after the initiation of contrast injection. The full chest CT scan is acquired during patient breath-hold and requires around five seconds. Next, the CT table is moved back to the location of the DCE-CT scanning section, and as quickly as possible the scanner transitions back to DCE-CT scanning mode. At approximately 115 seconds after initial contrast injection, the 21st DCE-CT snapshot is acquired, and snapshots in this secondary DCE-CT scanning period are acquired every five seconds until five additional DCE-CT snapshots have been acquired (resulting in 25 total DCE-CT snapshots). The exact times associated with snapshot acquisition are imprinted in the DICOM headers of the reconstructed images.

The contrast media used in our study is iohexol (Omnipaque, 350 mg/mL iodine concentration; Amersham Health, Princeton, NJ). The amount of contrast used in a DCE-CT scan varies from study to study, though recommendations are typically around 50 mL of iodinated contrast media

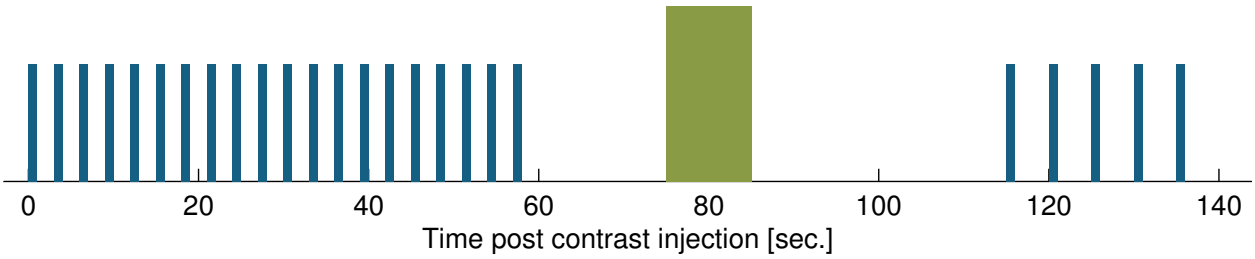


Figure 6.1: Graphical summary of the DCE-CT protocol used in this study. Blue lines indicate the DCE-CT imaging snapshots, and the green bar indicates the standard clinical contrast-enhanced chest CT.

injected at around 4 mL/second [77, 124]. However, in our study, the standard amount of contrast for a full chest CT was necessary because of the chest CT component. Since the entire premise of our protocol design was to "piggyback" on an otherwise necessary standard chest CT scan, the patient scans in our study were all performed with 90 mL of contrast (90 mL is on the low end of the standard range of contrast administrations for full chest CT scans). To somewhat compensate for the increased amount of contrast, injection was performed at an increased 6 mL/second. The contrast injection was immediately and automatically followed by a 30 mL "chaser" of saline flush, which ensured that the entire bolus of contrast had entered into the circulation of the patient.

6.2.3 DCE-CT Imaging Parameters

In order to select the imaging parameters for the DCE-CT snapshots, previous DCE-CT studies were investigated. In the study by Meijerink *et al*. [80], scanning parameters were a tube voltage of 80 kVp with a current of 120 mAs, acquiring snapshots every second for 30 seconds (or for the complete duration of patient breath-hold), leading to two adjacent reconstructed slices of 10 mm each. Two additional protocols from lung nodule studies were compared. The first, by Ma *et al*. [75], used 120 kVp and 60 mAs, acquiring snapshots approximately every half-second for 40 seconds and reconstructing four adjacent slices of 5 mm each. The second study, by Li *et al*. [83], used 120 kVp and 100 mAs, acquiring snapshots every 5 seconds for 60 seconds and re-

constructing 5 mm slices. From these studies, and the guidance offered in other DCE-CT literature [77, 78], the imaging protocol parameters selected for this study were 120 kVp with 100 mAs, acquiring snapshots every 3 seconds for 60 seconds during the first 20 snapshots, then every 5 seconds for the last group of five snapshots. Reconstructions were set for 3 mm slices (resulting in 18 axial slices in each DCE-CT snapshot) using the standard kernel, without edge enhancement. The choice of 120 kVp was a rational choice for thoracic imaging due to the physical size and attenuation of the human thorax, and 100 mAs offered a good balance between image noise and patient dose. Effective patient dose in CT scales approximately linearly with reported dose-length product (DLP) [125–127], and phantom studies during the planning phase indicated that the effective dose from the DCE-CT scanning components contributed approximately 1.5 times the dose of the standard chest CT scan. Finally, based on an analysis of contrast uptake curves in the aforementioned lung DCE-CT studies, a temporal sampling of three seconds was deemed sufficient to capture first-pass hemodynamic uptake curves.

6.3 Image Processing

After DCE-CT scan acquisition, the DCE-CT series are reconstructed using 3-mm axial slices (for analysis purposes) and 1-mm axial slices (for visualization purposes). Each DCE-CT series consists of 25 snapshots, each snapshot containing 18 3-mm sections, resulting in 450 axial sections for the thicker reconstruction and 1375 axial sections for the thinner reconstruction (the full axial span of the DCE-CT series is 55 mm). The images were pushed from the CT workstation to the network picture archiving and communication system (PACS) and were stored as 16-bit integer images, with data restricted to only the first 12 bits. The image headers define the rescaling intercept and slope to be used with the pixel value data, but the native headers are incorrectly assigned as the inverse of the correct rescale slope; if the slope is not manually corrected, the image values will not be true HU values, and the resulting analysis will not be indicative of true hemodynamic parameters.

Before proceeding further, the DCE-CT series are visualized across both spatial and temporal dimensions. Using a simple script written for the in-house software package Abras, the user can interactively visualize a single axial section across time or can visualize the entire spatial image stack for a single snapshot. These visualizations aide in the identification of gross motion artifacts, as well as granting an initial impression of contrast uptake and washout through the different anatomical structures of interest. Because the patients continuously respire and have continuously beating hearts, some motion is expected in the DCE-CT images, and this initial visualization allows the user to identify any locations where motion (especially axial motion across the superior or inferior boundaries of the spatial image stack) may lead to unreliable estimates of contrast uptake curves.

6.3.1 Image Registration: Initial Efforts

In order to remove the effects of patient motion (both internal and external), deformable image registration was used to deform each individual DCE-CT snapshot in a DCE-CT scan to a common reference frame. Image registration is a common tool often used to facilitate analysis of serial CT scans of the same patient or 4DCT scans (as in the current study) [128–131]. Registrations fall into three general categories: rigid, affine, and deformable registrations. In a rigid registration, the moving image volume is allowed to rotate and translate (and sometimes scale) to match a target image volume. In an affine registration, the moving image volume is further allowed to shear to match the target image volume. In a deformable registration, the vectors describing the warp field between the moving image and the target image are approximately spatially independent (only true to the extent of regularization of the deformation field).

The registration task in this study is to register all the DCE-CT snapshots of a scan to a common reference point. The image coordinate systems are identical (same scanner, same table position, same day) and the patient does not physically change much in terms of scaling or shear, and therefore a rigid or affine registration is mostly unnecessary. However, a deformable registration is

necessary to match DCE-CT snapshots, since while the patient bones and muscles are not moving (much), the lungs and heart are continuously moving. Therefore, the posterior surface of the patient is nearly fixed while the anterior surface is rising and falling, and the interior anatomy is expanding and contracting with the respiratory cycle. In order to focus the registration task on the patient only (and not include the CT table), the thoracic segmentation described in Chapter 4 was applied to all DCE-CT images before image registration.

For this study, the "target" image was always selected as snapshot number 20 (i.e., the last snapshot of the initial DCE-CT series) instead of "daisy-chaining" the successive deformation fields together (e.g., $1 \rightarrow 20$ directly rather than a composite deformation vector field generated from $1 \rightarrow 2$, $2 \rightarrow 3$,..., $19 \rightarrow 20$). Snapshot 20 is approximately half-way through the total scan timeline, and the enhancement of the structures in the image volume are at a qualitatively "moderate" level. Use of the first snapshot as the target reference frame would have resulted in registration challenges when the contrast media washes into the heart and greatly enhances the mediastinal structures (e.g., at peak enhancement of the mediastinal structures around the sixth snapshot, there is nothing in the first snapshot that these bright structures can "latch onto"). Use of a snapshot with highly concentrated contrast and bright structures would have led to registration challenges with the earliest and latest snapshots, when the enhancement is diminished. Empirical exploration of the registration process identified the 20th snapshot as a reasonable compromise for the target image reference frame.

There are many options available for deformable registration algorithms, but probably the most popular such algorithm is called "demons" [132] (after the thermodynamic thought experiment postulated by James Clerk Maxwell) and was first implemented by J. P. Thirion [133]. The demons registration algorithm relies on an optical flow technique to match image value level sets. A software package called "Plastimatch" was initially used for deformable demons registration in this study, since the package had been heavily parallelized and was freely available from the developers [134]. The Plastimatch registration was implemented first using an affine registration, then

a multi-resolution-stage demons registration. At the first stage, the volumetric image block was down-sampled by eight times for in-plane resolution, and axial resolution was made as isotropic as possible to the new down-sampled resolution. The stages progressed to four-times down-sampling, two-times down-sampling, and finally native resolution registration. This resolution "ladder" allows for the moving image block to be matched to the target on a very coarse level initially, with refinements for the deformation vector field happening at each successive rung on the ladder, which allows for an increase in computational efficiency. The metric used in registration optimization was the global mean squared error (MSE) between matched pixel values, which is the default (and only) choice for demons registration in Plastimatch.

The registration of one moving image block to the 20th snapshot target image block required approximately two minutes on a desktop quad-core 64-bit PC. The process was parallelized and utilized nearly 100% of all four CPU cores for the duration of the registration. Therefore, the total time required to register an entire DCE-CT scan (with 25 DCE-CT snapshots) was around 50 minutes.

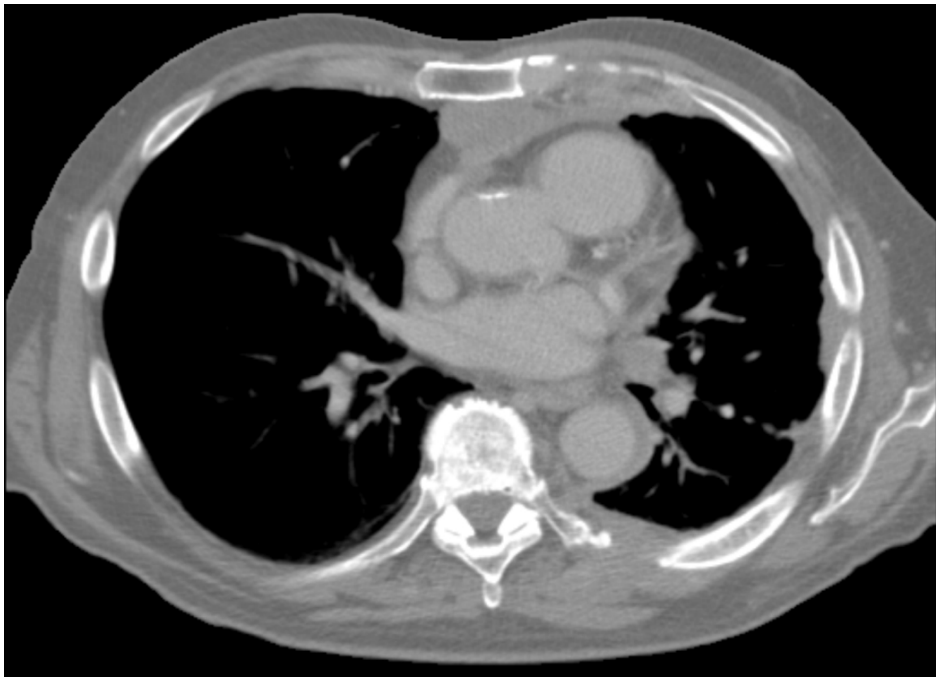
Figure 6.2 shows an axial image from the 20th DCE-CT snapshot of an example patient. Figure 6.3 shows the matched axial section from the Plastimatch deformed 8th snapshot, where is can be seen that the concentration of contrast is fairly high in the vascular structures. The images are spatially well-matched, especially the portion of mesothelioma immediately posterior to the patient's sternum, but the demons registration process introduces image artifacts in structures with changing HU values due to the use of the MSE metric. Since the changing concentration of iodinated contrast in the patient's vasculature will alter the underlying attenuation of those structures, it is expected that they will change HU value on the CT images. However, the MSE metric penalizes differences in HU value quadratically, leading to artifacts especially at the boundaries between structures of different HU value. For instance, Figure 6.3 exhibits deformation artifacts around the border of the bright mediastinal structures and the pulmonary arteries, especially in the patient's right lung. The Plastimatch registrations very effectively "fix" structures in space across time, but

the appearance of the aforementioned artifacts is ubiquitous across all snapshots.

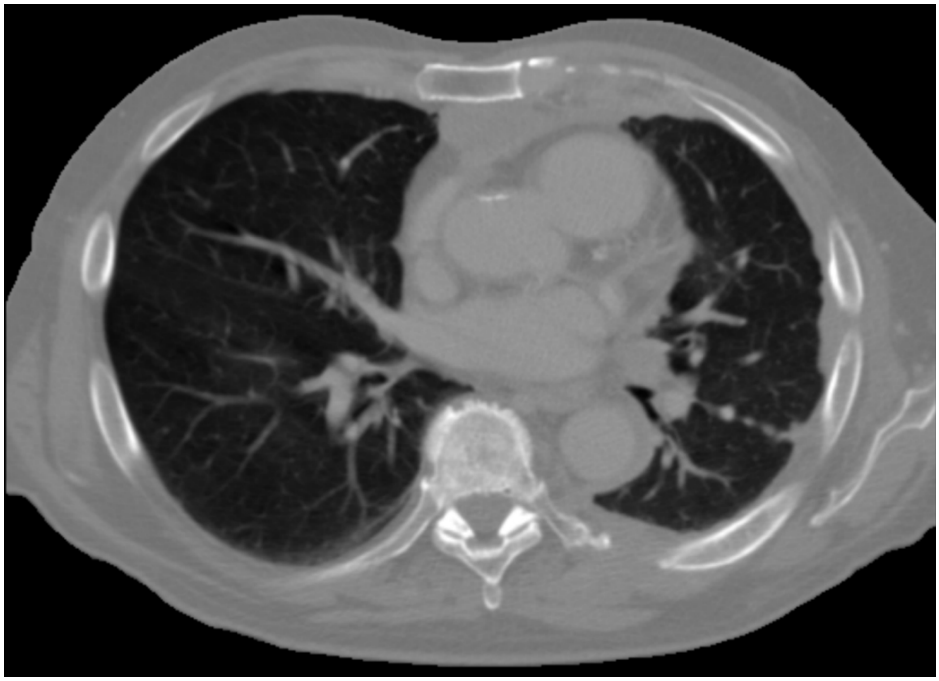
6.3.2 Image Registration: Updated Efforts

After the Evaluation of Methods for Pulmonary Image Registration 2010 (EMPIRE10) challenge results became public [135], we became interested in the deformable registration algorithm that performed best, offered by a software package called Advanced Neuroimaging Tools (ANTs) [136]. The primary deformable registration algorithm implemented in ANTs is a diffeomorphic symmetric normalization (SyN) transform. "Diffeomorphic" describes certain properties of the resulting vector deformation field, guaranteeing that the field will be mathematically "well-behaved" with properties such as differentiability and invertibility. The SyN transform used is a symmetric technique to derive the underlying vector deformation field of the diffeomorphism by starting at the target image and the moving image and working toward the middle. The image is deformed with no smoothing applied to the deformation field, as is typical for a fully deformable registration technique (e.g., demons). The parameters for the derivation of the diffeomorphism were set to nearly default values as recommended by the program developers [136], with slight tweaks to allow for larger deformations when necessary.

There are more deformation metrics from which to choose for ANTs registration than for Plastimatch demons registration. We found that the recommendations of the developers provided the best performance for our registration task (binned mutual information for the affine component and normalized cross-correlation in the default local image neighborhood for the deformable SyN registration component) [136]. Using cross-correlation as the metric for diffeomorphism optimization has obvious advantages for our registration task; since HU values are expected to change, MSE metrics will perform poorly when faced with different contrast concentrations across DCE-CT snapshots (as indicated previously). However, at least in small location neighborhoods on the order of 5x5x5 voxels, HU values can still be expected to correlate (i.e., bright with bright, dark with dark) between snapshots even if they are not identical. That is, the global MSE metric enforces

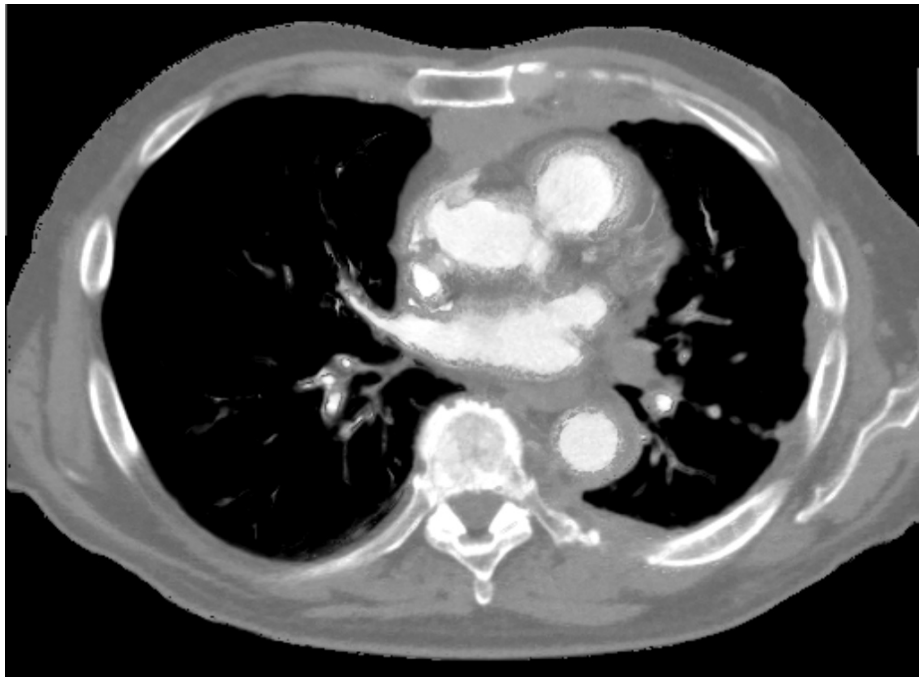


(a)

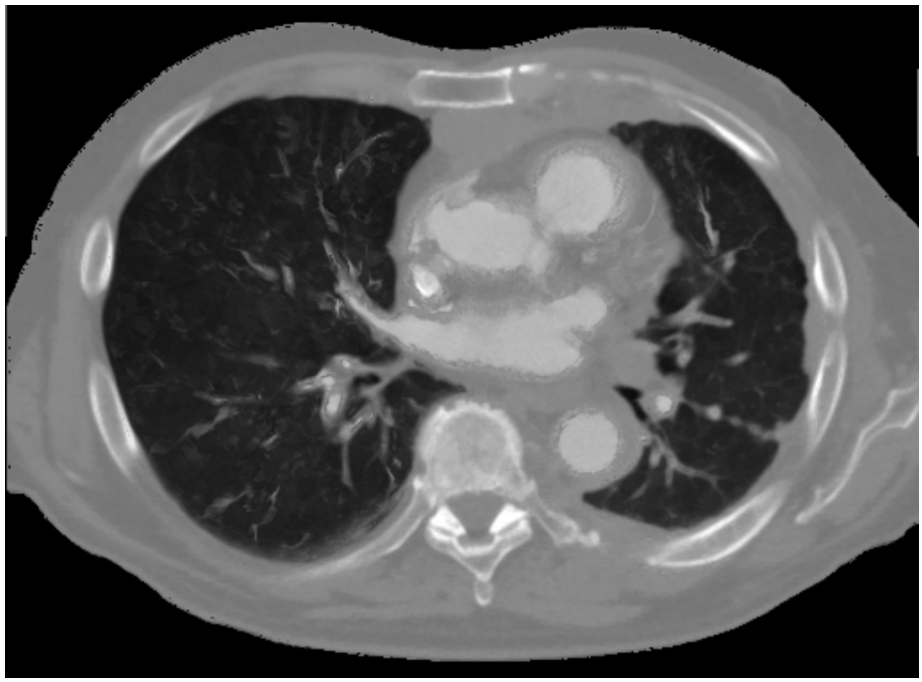


(b)

Figure 6.2: Axial image (undeformed) from the 20th DCE-CT snapshot of an example patient scan. (6.2a), the grayscale window level is set to 0 HU, and the window width is set to 1000 HU. (6.2b), the window level/width is set to 0 HU/2000 HU.



(a)



(b)

Figure 6.3: Plastimatch-deformed axial image from the 8th DCE-CT snapshot of the same patient scan shown in Figure 6.2. (6.3a), grayscale window level/width is set to 0 HU/1000 HU. (6.3b), the window level/width is set to 0 HU/2000 HU.

an identity relationship between HU values in the target and moving images over the *entire* image blocks, while the cross-correlation metric only enforces that the relationship between HU values between the two image blocks be linear in a local neighborhood. Cross-correlation is an inherently better metric to use for our registration task, and the results shown below for the ANTs registration illustrate that fact. Finally, the ANTs registrations were all performed with a multi-resolution-stage ladder similar to that of the Plastimatch registrations, since these multi-scale registrations are fairly common in the image registration literature.

One downside of the ANTs registration software is that it is not effectively parallelized, using only one CPU core for the vast majority of computation. The calculation of the cross-correlation metric is also fairly time consuming; the time required to register one DCE-CT snapshot to its corresponding target image is approximately 45 minutes on a desktop 64-bit PC. However, the registration process can be entirely parallelized using The University of Chicago's Scientific Image Reconstruction and Analysis Facility (SIRAF), since for a DCE-CT scan as a whole, there are 24 independent registration tasks that can occur simultaneously (snapshots $1, 2, \ldots, 19, 21, \ldots, 25$ all need to be registered to snapshot 20). Since there are hundreds of cores on the SIRAF cluster, even assigning two cores per registration task for computational overhead results in all 24 registration tasks occurring simultaneously, and the registration of the entire DCE-CT scan is therefore completed in approximately 45 minutes.

Figure 6.4 shows the axial section matched to Figure 6.2 from the ANTs deformed 8th snapshot. These images highlight the improvement in subjective image appearance. Anatomic structures are still well-matched spatially, but the deformed image is free of the artifacts exhibited by the demons registration (see Figure 6.3). When the unique DCE-CT snapshots of a single demons-registered axial section are viewed across time (i.e., as a "cine loop"), individual structures are very effectively "pinned down" in that they are all well-registered to the same target coordinate system. However, the appearance of any one snapshot will include demons artifacts. With the ANTs registrations, structures are well-matched, but boundaries may exhibit a bit of "wobble" across the

temporal dimension on the order of one or two voxels in-plane. That is, the same cine loop of a single ANTs-registered axial section will be largely free of deformation artifacts but will reveal very slight mis-registrations across time, especially at the boundary between adjacent structures. Since uptake curves will be averaged over spatial structures, however, it is easy to avoid the artifacts in these boundary regions by simply padding the contours of spatial structures by one or two voxel widths. The same contour padding would also be necessary with the demons registrations due to the artifacts in regions of high HU-value gradients. Because of the qualitatively improved image appearance with the ANTs registrations, ANTs became the registration algorithm of choice for the remainder of this DCE-CT study.

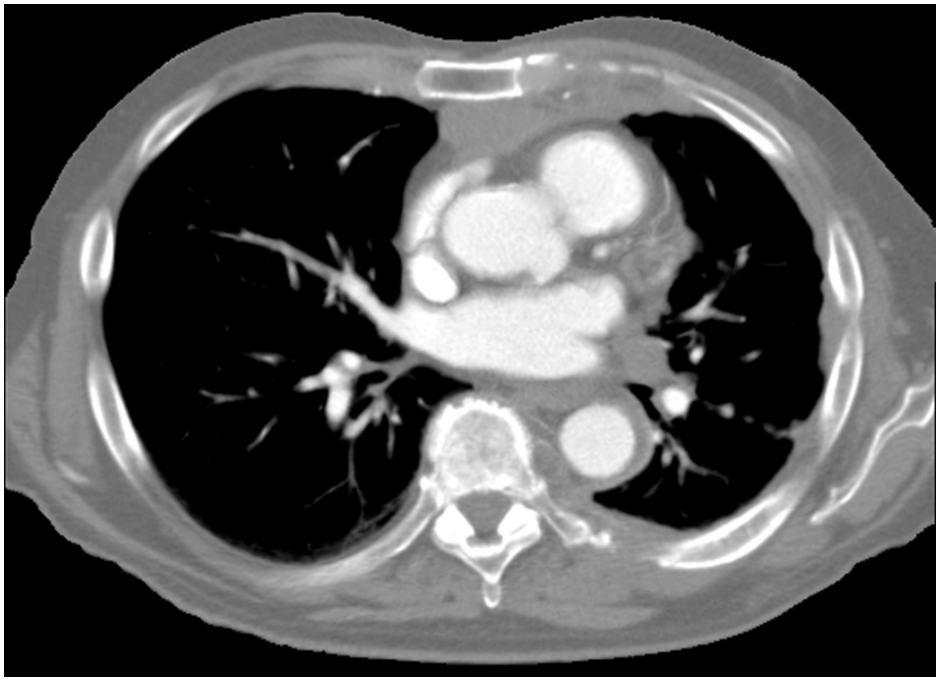
6.4 Dynamic Analysis

6.4.1 DCE-CT Parameter Calculation

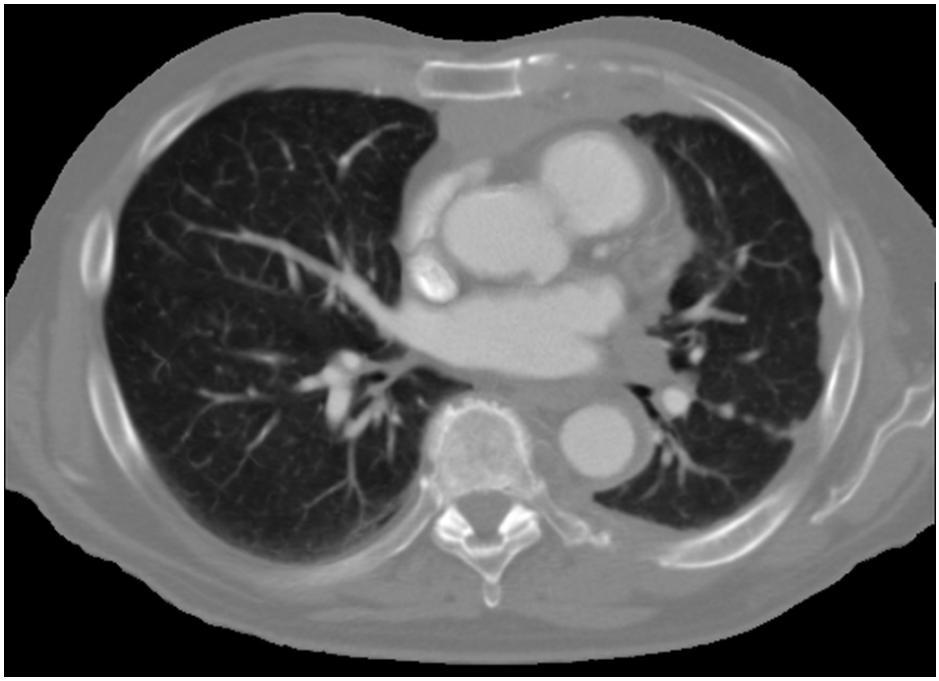
After image registration has been performed, the HU value of an individual voxel can be tracked over time, and the extent of HU value increase over baseline will be proportional to the concentration of contrast media in the voxel. One uptake curve is obtained for every voxel in the three-dimensional image, and various hemodynamic parameters can be derived from these curves. The first-pass kinetic parameters used in this research include perfusion, tissue peak enhancement, calculated blood volume, time to peak enhancement, and mean transit time [77, 79].

Perfusion measures the flow rate through the vasculature of the structure and is an oncologic marker for tumor vascularity and grade. Perfusion (along with the other parameters) is calculated according to the "slope method" of deconvolution analysis and is calculated as the ratio of the maximum slope in the tissue uptake curve to the peak arterial enhancement, or

$$\text{Perfusion} = \frac{\max\left(\frac{\mathrm{d}HU(t)_{tissue}}{\mathrm{d}t}\right)}{\max\left(HU(t)_{arterial} - HU(0)_{arterial}\right)}. \tag{6.1}$$



(a)



(b)

Figure 6.4: ANTs-deformed axial image from the 8th DCE-CT snapshot of the same patient scan shown in Figure 6.2. (6.4a), grayscale window level/width is set to 0 HU/1000 HU. (6.4b), the window level/width is set to 0 HU/2000 HU.

This formulation of perfusion is the slope method of the original Mullani-Gould approach [137], and the derivation of the formula is found in the Appendix of reference [79]; the method has been validated in other organ systems [138, 139]. For the discrete data in this study, the maximum slope in the numerator of equation 6.1 is calculated using the running difference between values in the uptake curve, divided by the temporal spacing of points on the uptake curve. The denominator is calculated as the maximum enhancement of the average uptake in an arterial region of interest over its baseline value.

Peak enhancement is the maximum increase in tissue density over baseline, or

$$\text{Peak Enhancement} = \max\left(HU(t)_{tissue} - HU(0)_{tissue}\right), \tag{6.2}$$

and is an oncologic marker of tissue blood volume. The marker commonly called "blood volume" is simply a normalized version of peak enhancement and is calculated by the slope method as

$$\text{Blood Volume} = \frac{\max\left(HU(t)_{tissue} - HU(0)_{tissue}\right)}{\max\left(HU(t)_{arterial} - HU(0)_{arterial}\right)}. \tag{6.3}$$

This formulation of blood volume was presented by Koenig *et al.* [140] and is an approximation of blood volume as originally presented by Axel *et al.* [141].

The time to peak enhancement is a measure of the time for contrast to flow from a major arterial vessel to the tissue, or

$$\text{Time To Peak} = \arg\max\left(HU(t)_{tissue}\right) - t_{\text{arterial arrival}}, \tag{6.4}$$

and is an oncologic marker of blood pressure (where $t_{\text{arterial arrival}}$ is the time of initial contrast enhancement in the arterial input).

Mean transit time is the time taken for contrast to travel from artery to vein, calculated as the

full-width at half-maximum (FWHM) of the tissue uptake curve over baseline, or

$$\text{Mean Transit Time} = \text{FWHM}\left(HU(t)_{tissue} - HU(0)_{tissue}\right), \tag{6.5}$$

and is an oncologic marker of local perfusion pressure.

6.4.2 Region of Interest Identification

Many of the above DCE-CT parameters require the definition of an arterial input function (AIF). In our patient cohort, the pulmonary arteries (PA) were used whenever possible; contours were placed in the PA trunk at the level of the PA bifurcation when the structures were in the reconstructed field of view. For other patients whose PA trunk was not visible in the axial DCE-CT section, the PA main branch corresponding to the diseased hemithorax was used as the AIF. Contouring for all regions of interest, including the AIF, was performed by the author using our in-house software Abras. The average HU value of the voxels inside the arterial input contour was used as the value of the AIF for the time associated with a given DCE-CT snapshot.

Other regions of interest were contoured for each patient. Spatially isolated foci of disease were contoured separately, and contours were allowed to span multiple axial slices (volumetric contours). Because the scans had been registered to a common reference coordinate system (the 20th DCE-CT snapshot), the contours only needed to be drawn on a single snapshot of the observer's choosing (e.g., the snapshot that best highlighted the difference between the structure of interest and the surrounding structures). Matched ROIs were independently contoured on the second DCE-CT scan for each patient with reference to the first DCE-CT scan. All contours were verified to include the region of interest for *all* DCE-CT snapshots by overlaying the contour across the temporal images for a given axial section. Because of the "wobble" mentioned earlier for the ANTs registration algorithm, contours were necessarily drawn a few voxel-widths interior to the boundary of the disease foci; in this way, the ROI contours included the distinct disease foci across

all DCE-CT snapshots. A median of four distinct volumetric regions of interest were contoured for each patient, each corresponding to a distinct focus of disease.

After registration, the uptake curve of each voxel inside the patient's thorax is calculated (i.e., the HU-value trajectory across time for the 25 DCE-CT snapshots). Individual voxel uptake curves are subject to image noise, leading to variability in the DCE-CT parameters calculated by the slope method shown above. Therefore, before the voxel-by-voxel DCE-CT data maps can be displayed as images, smoothing was applied to individual voxel uptake curves using the "csaps" cubic smoothing spline function in MATLAB, an extension of the "smooth" function from FORTRAN. The default value of the smoothing parameter was used for data points with a separation of approximately six seconds, since it was reasonable to believe that while uptake curve changes on the three-second sampling scale might be affected by noise, changes at half the sampling rate may be more robust to image noise. It should be noted, however, that these smoothed uptake curves are *only* used to visualize the DCE-CT data maps.

6.4.3 DCE-CT Parameter Values

DCE-CT parameters can also be summarized from the average uptake curves of the individual ROIs. That is, the average HU value of all the voxels inside a disease ROI can be extracted from each DCE-CT snapshot using the observer-provided contours, and this single "average uptake curve" can be used to calculate the DCE-CT parameters for the individual ROI. Average uptake curves are shown for an example patient scan (patient #2, scan #2) in Figure 6.5. The effect of spatially averaging the HU values inside a contour at each DCE-CT snapshot is sufficient to not require any further smoothing of the average uptake curves. The DCE-CT parameter values reported below are all derived from average uptake curves.

The implementation of the DCE-CT analysis software in MATLAB is flexible; it can handle DCE-CT scans with arbitrary time spacing and arbitrary definitions of regions of interest. Many implementations of DCE-CT or "perfusion CT" analysis software, especially commercial, limit

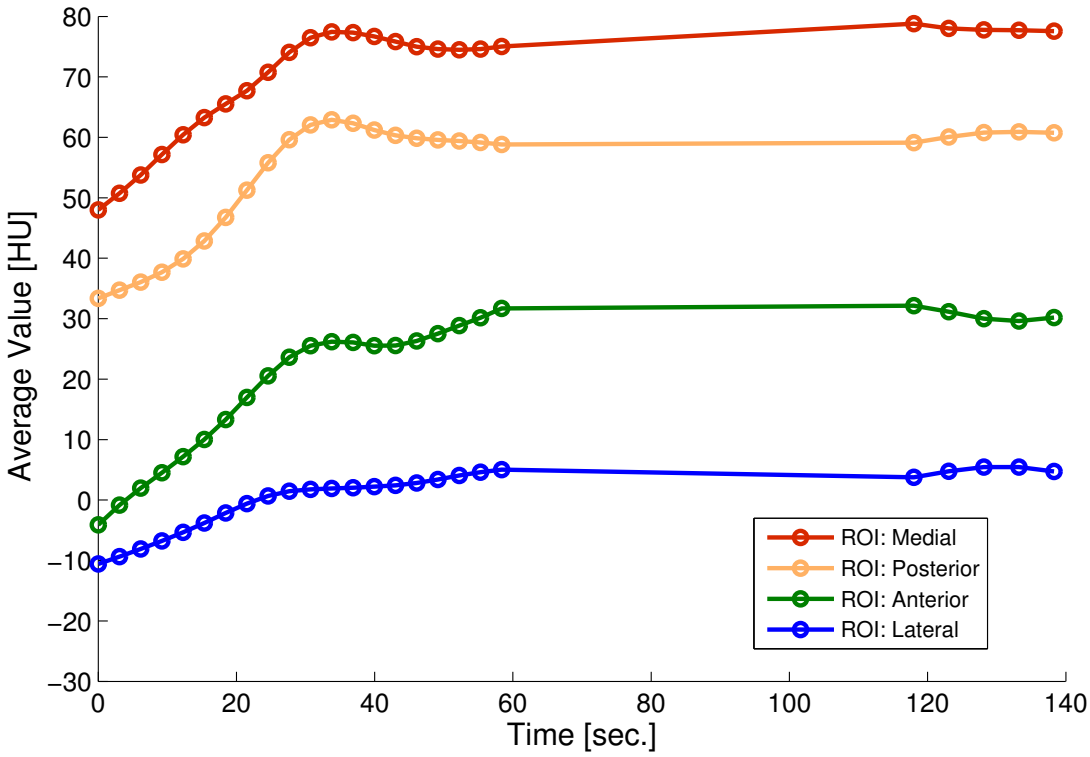


Figure 6.5: Average uptake curves from an example patient's tumor regions of interest across all 25 DCE-CT snapshots. The four distinct foci of disease were each identified by the author using volumetric contours.

the user to identifying regions of interest as simple shapes (ovals or boxes) in single axial sections. The code written specifically for this study allows contours placed on different and potentially non-sequential axial sections to be linked, and the average uptake curves for those regions of interest are simply derived from any voxels inside the mask formed by the volumetric contours. Volumetric contouring is especially useful for more "rind-like" tumor ROIs, where the thickness of disease in any one axial section can be rather thin, but many axial sections contain the same spatial region of interest (perhaps "volume of interest"). The voxel count for these ROIs is increased substantially (and the noise in the average uptake curve thereby reduced) through the use of volumetric contours rather than single-section contours. The code for this study is relatively fast, calculating all DCE-CT parameters for the entire collection of voxels in approximately two minutes on a standard desktop PC. However, even with parallelization across all four cores of a quad-core computer, the spline smoothing operation required for image visualization purposes mentioned above can take approximately one hour.

Table 6.2 summarizes the DCE-CT parameters for the ROIs from patients on observation. Since there are seven patients on observation, there are 14 individual scans included in these summaries (among the seven patients, there are 19 unique ROIs). Some values of mean transit time (from equation 6.5) are not calculable, since there is no guarantee that the average uptake curve will fall to its half-maximum value within the allotted DCE-CT imaging time. Therefore, all that can be said for the full-width at half-maximum of uptake over baseline for these ROIs is that it exceeds the total observed imaging duration. Table 6.3 shows the same DCE-CT parameter summaries for the ROIs from patients being treated. There are six patients "on treatment," and therefore there are 12 scans included in these summaries (among the six patients, there are 24 unique ROIs).

The parameter "time to peak" (TTP) is interesting to study at a single time point, while the other DCE-CT parameters are perhaps more interesting when comparing DCE-CT scans across time (i.e., using the DCE-CT parameters as imaging biomarkers of response). In order to maximize the enhancement of MPM on a single standard chest CT scan, the delay between contrast injection and

DCE-CT Parameter	Mean Value (first scan)	St. Dev. (first scan)	95% CI (first scan)	Mean Value (second scan)	St. Dev. (second scan)	95% CI (second scan)
Perfusion (equation 6.1) [1/sec]	0.00659	0.00424	[0.00161, 0.0162]	0.00819	0.00807	[0.00198, 0.0291]
Peak Enhancement (equation 6.2) [HU]	28.7	23.8	[1.08, 82.1]	22.5	22.1	[0.638, 70.9]
Blood Volume (equation 6.3)	0.0479	0.0335	[0.00193, 0.122]	0.0473	0.0501	[0.00113, 0.160]
Time To Peak (equation 6.4) [seconds]	81.0	43.7	[6.76, 121]	73.3	51.9	[2.76, 124]
Mean Transit Time (equation 6.5) [seconds]	47.5	43.1	[3.94, 115]	31.1	35.1	[3.49, 87.1]

Table 6.2: Summary of DCE-CT parameters from ROI average uptake curves taken from patients on observation. The 95% confidence interval (CI) is calculated as the 2.5% and 97.5% percentiles. For the first DCE-CT scan, 47.4% of mean transit time values are not calculable, and for the second DCE-CT scan, 63.2% are not calculable.

DCE-CT Parameter	Mean Value (first scan)	St. Dev. (first scan)	95% CI (first scan)	Mean Value (second scan)	St. Dev. (second scan)	95% CI (second scan)
Perfusion (equation 6.1) [$^{1}/_{\text{sec}}$]	0.00900	0.00439	[0.00341, 0.0193]	0.00808	0.00408	[0.00264, 0.0170]
Peak Enhancement (equation 6.2) [HU]	31.6	18.0	[9.42, 79.0]	31.7	19.1	[7.38, 78.3]
Blood Volume (equation 6.3)	0.0707	0.0419	[0.0222, 0.179]	0.0667	0.0418	[0.0167, 0.170]
Time To Peak (equation 6.4) [seconds]	80.3	50.9	[12.3, 134]	69.6	44.6	[14.1, 125]
Mean Transit Time (equation 6.5) [seconds]	40.5	45.7	[4.28, 110]	60.3	40.7	[10.6, 103]

Table 6.3: Summary of DCE-CT parameters from ROI average uptake curves taken from patients on treatment. The 95% confidence interval (CI) is calculated as the 2.5% and 97.5% percentiles. For the first DCE-CT scan, 79.2% of mean transit time values are not calculable, and for the second DCE-CT scan, 70.8% are not calculable.

scan acquisition should be set to the time required for peak enhancement, quantified in this study as TTP. From the TTP data, we have an overall average TTP value of 76.2 seconds when using all patient ROIs. The average value of $t_{\text{arterial arrival}}$ from equation 6.4 is 6.3 seconds, indicating that the average peak in disease ROI uptake occurs 82.5 seconds after initiation of contrast injection. However, for a standard chest CT scan with contrast, it is typical to acquire the scan approximately 70 seconds after initiation of contrast injection. Therefore, if clinicians wanted to acquire a chest CT specifically to maximally enhance tumor tissue for patients with MPM, it may be prudent to wait additional time before acquiring the CT scan post contrast injection (for the contrast injection parameters given in this study). However, any additional delay may negatively impact the visibility of mediastinal structures and pulmonary vasculature.

Images of the DCE-CT parameter data maps for an example scan (patient #2, scan #2) are shown in Figures 6.7 through 6.11. Figure 6.6 shows the temporal maximum intensity projection (tMIP), where the maximum HU value across time for each voxel is assigned to the spatial location of that voxel. Images of tMIP maps are useful for visual and anatomic reference, since they combine data across all DCE-CT snapshots. Figure 6.7 shows the perfusion map, Figure 6.8 shows the peak enhancement map, Figure 6.9 shows the blood volume map, Figure 6.10 shows the time to peak map, and Figure 6.11 shows the mean transit time map. The main tumor region on these maps is shown with an outline and an arrow. The tMIP image reveals that the main posterior mass adjacent to the vertebral body contains some low-uptake voxels (more anterior) and some higher-uptake voxels (more posterior), though the first DCE-CT snapshot image of this mass is nearly spatially uniform. The posterior higher uptake region is also reflected in the perfusion, peak enhancement, and blood volume maps. In these figures, no effort was made to mask the uptake in the lungs, which have very high perfusion properties due to the vascularity of the lung parenchyma and the extent of large vasculature in the lungs.

Figure 6.6: Example axial section of the temporal maximum intensity projection (tMIP) for patient #2, scan #2. The units of the image values are HU. The ROI indicated by the green arrow is the "posterior" ROI shown in Figure 6.5.

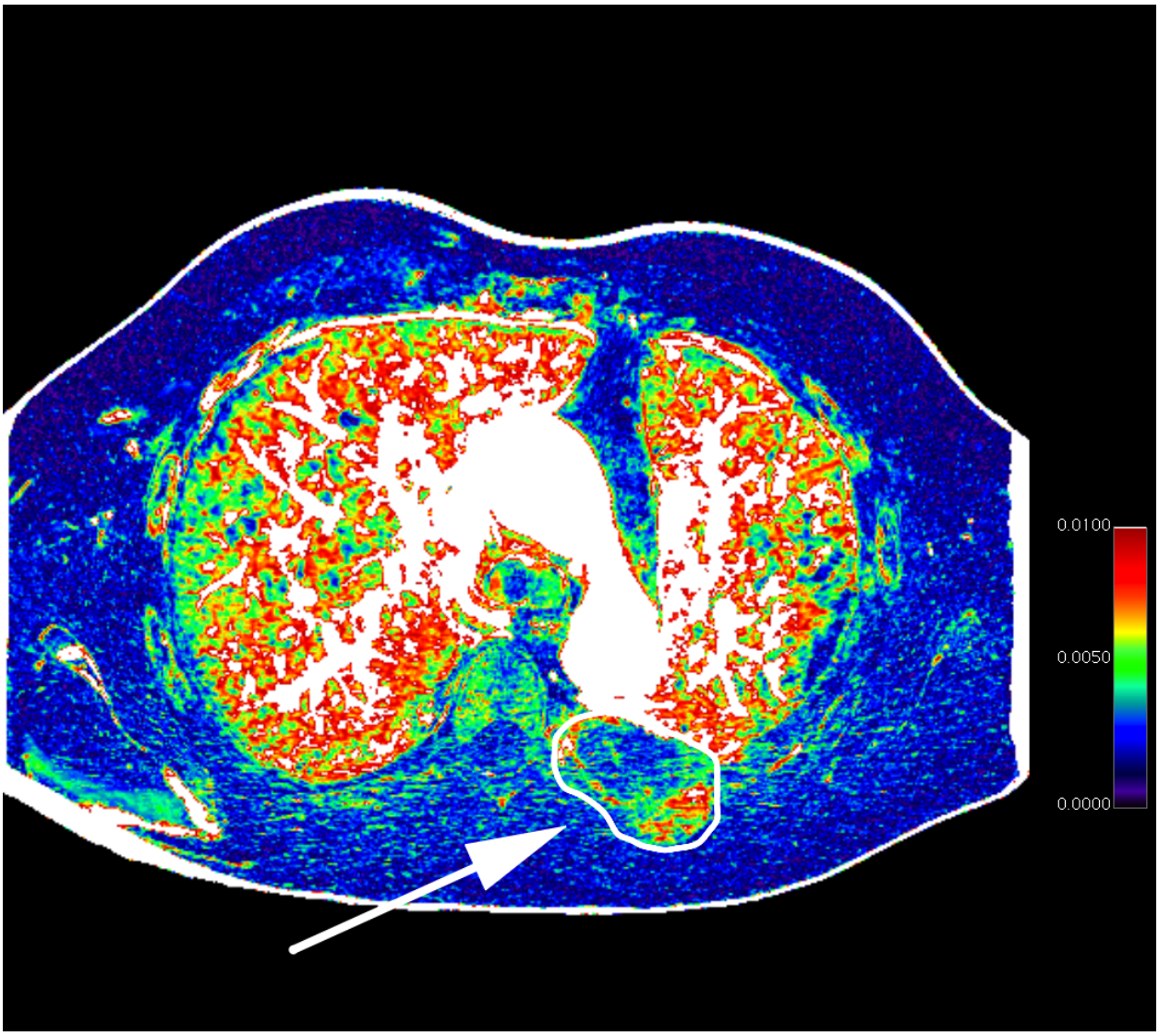


Figure 6.7: Example axial section of the perfusion map for patient #2, scan #2. The units of the image values are $^1/_{\text{sec}}$. The ROI indicated by the white arrow is the "posterior" ROI shown in Figure 6.5 and has an average uptake curve perfusion value of 0.0038 $^1/_{\text{sec}}$.

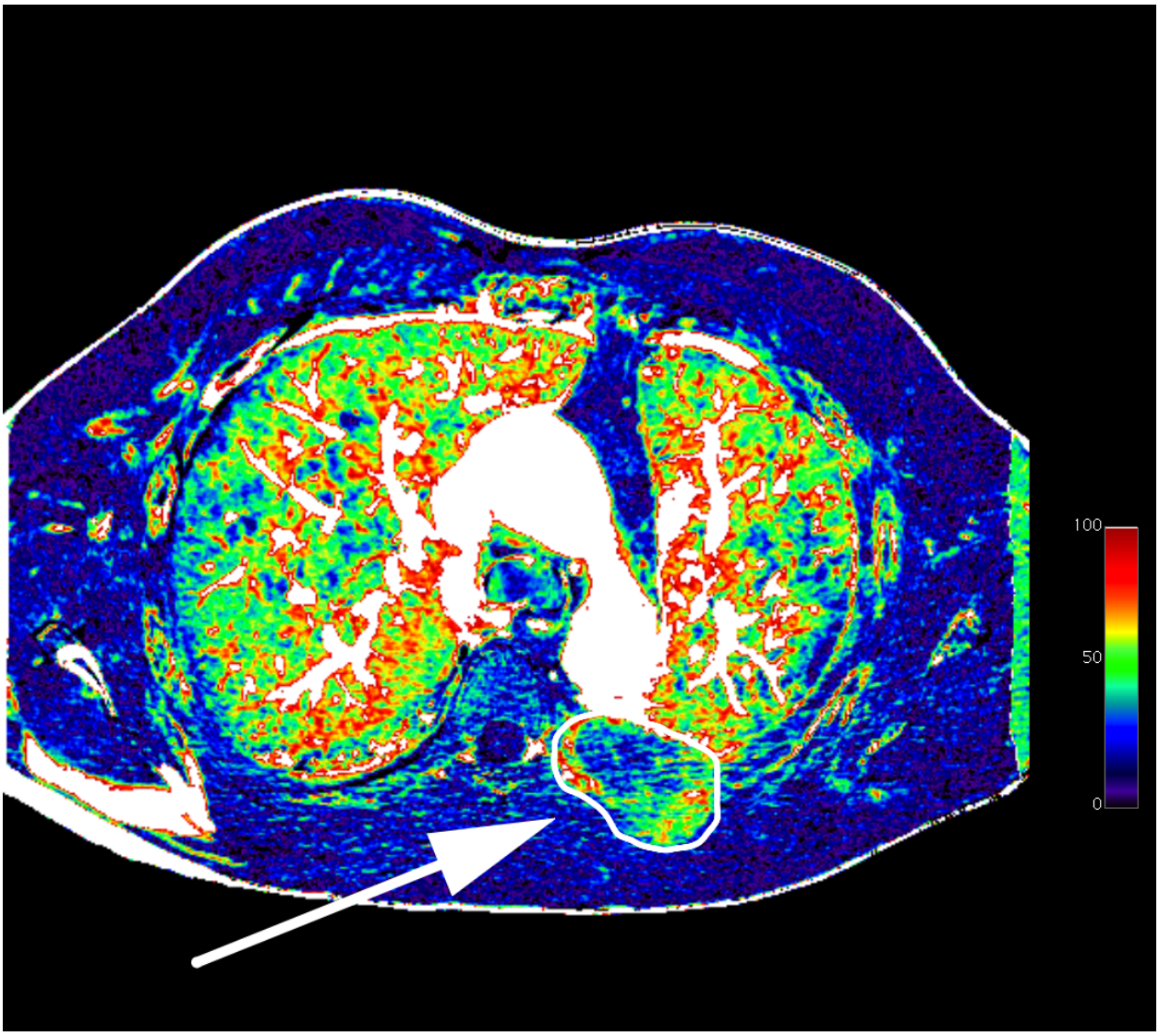


Figure 6.8: Example axial section of the peak enhancement map for patient #2, scan #2. The units of the image values are HU. The ROI indicated by the white arrow is the "posterior" ROI shown in Figure 6.5 and has an average uptake curve peak enhancement value of 33.1 HU.

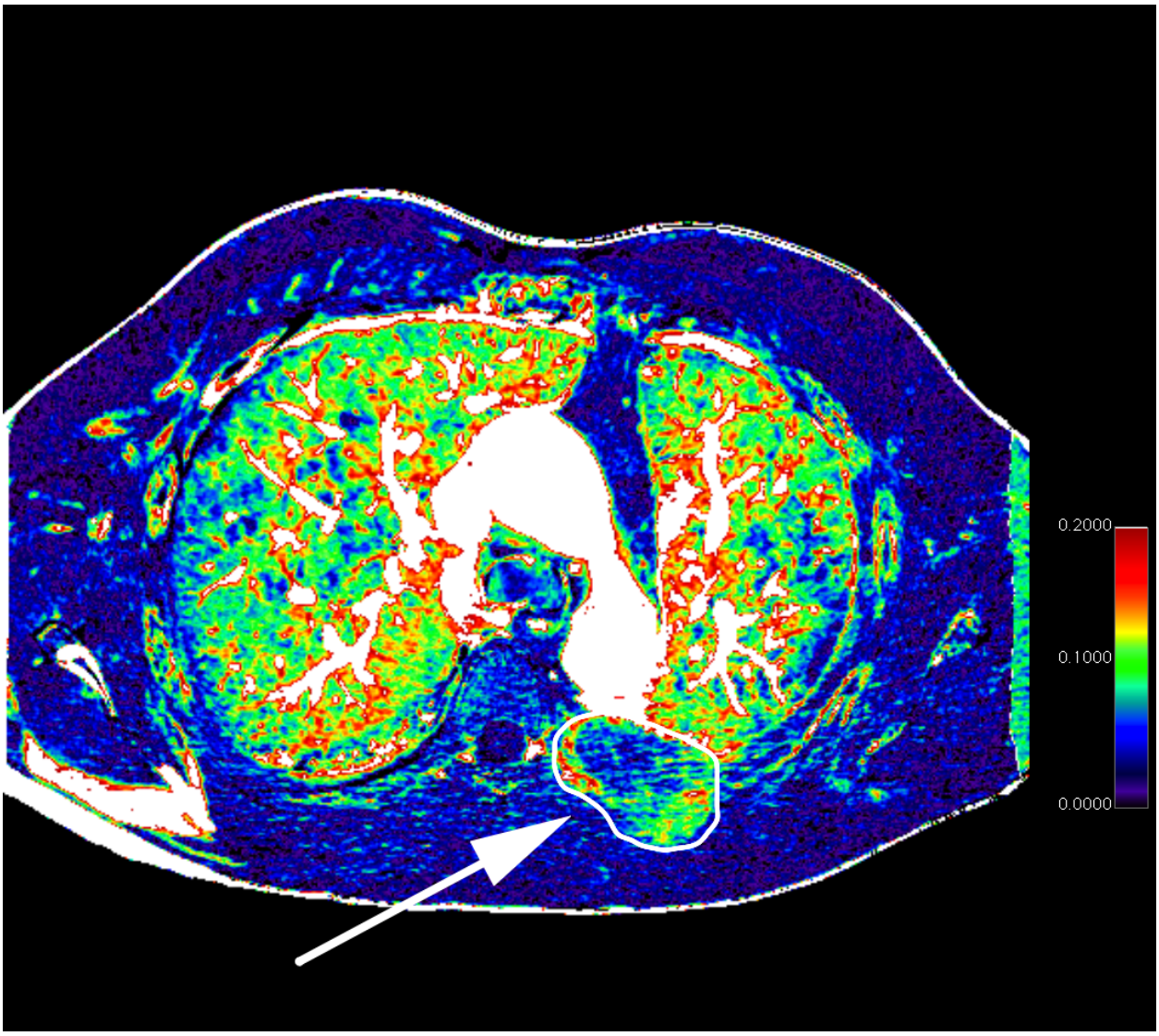


Figure 6.9: Example axial section of the blood volume map for patient #2, scan #2. The image values are unitless. The ROI indicated by the white arrow is the "posterior" ROI shown in Figure 6.5 and has an average uptake curve blood volume value of 0.061.

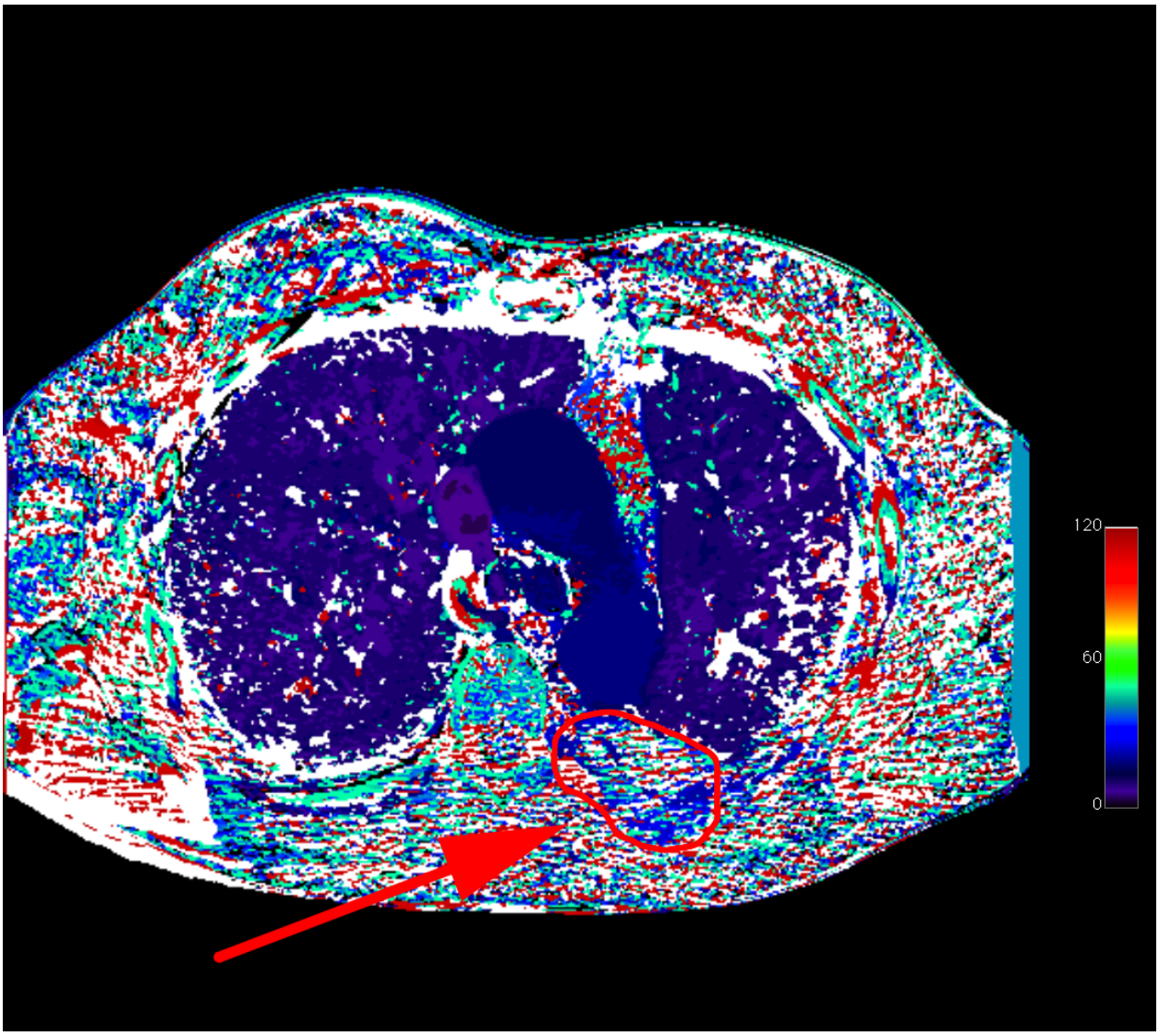


Figure 6.10: Example axial section of the time to peak (TTP) map for patient #2, scan #2. The units of the image values are seconds. The ROI indicated by the red arrow is the "posterior" ROI shown in Figure 6.5 and has an average uptake curve TTP value of 27.6 seconds.

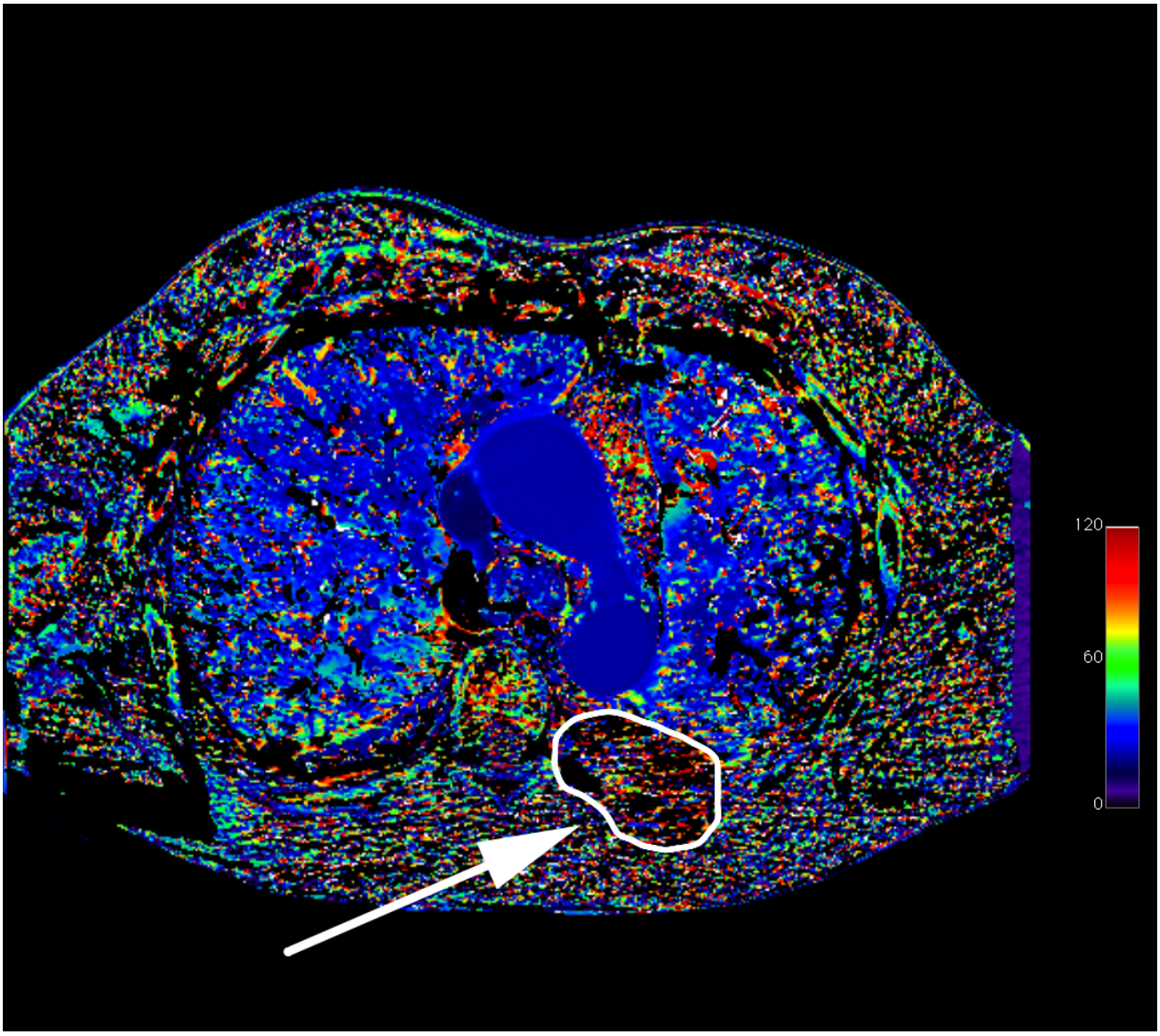


Figure 6.11: Example axial section of the mean transit time (MTT) map for patient #2, scan #2. The units of the image values are seconds. Voxels where the MTT value is not calculable (i.e., only known to be longer than the total duration of the DCE-CT scan) are shown as zeros. The ROI indicated by the white arrow is the "posterior" ROI shown in Figure 6.5 and the average uptake curve MTT value is not calculable.

6.5 Quantitative Response Assessment

6.5.1 Tumor Bulk

As mentioned previously, one component of the patient scan procedure was a full chest CT scan acquired at full inspiration, beginning approximately 75 seconds after the initiation of contrast injection. Tumor burden can be measured using these full chest CT scans for the purpose of investigating correlations between changes in tumor burden measurements and DCE-CT parameter measurements. Measurements of tumor burden are often used as tools for response assessment, but changes in tumor burden may not be directly related to changes in the hemodynamic properties of the disease. Therefore, we do not necessarily expect high correlations between changes in tumor burden and changes in DCE-CT parameters.

Tumor burden was quantified in the full chest CT scans using two distinct methods: first, the full volume of pleural disease was segmented using the semi-automated technique described in Chapter 4. Disease volumes from both scans per patient for all 13 patients who have currently undergone both DCE-CT scans in this study were quantified from these segmentations using the pixel-counting technique [94]. Second, linear measurements of tumor thickness were made on all scans according to the modified Response Evaluation Criteria In Solid Tumors (RECIST) technique [47]. A research radiologist experienced in thoracic CT and MPM measurements made six thickness measurements per patient scan (two measurements on each of three axial sections, per the modified RECIST protocol) with full access to both the first and second scan, per patient. The change in summed linear thickness and the change in volume of segmented pleural disease were recorded as measures of tumor "response."

Figure 6.12 shows the change in disease volume between the two scans as a function of the change in summed tumor thickness between the two scans. Compare this plot to Figure 5.4; in Chapter 5, it was shown that the rank correlation between changes in summed linear measurements and disease volume measurements was $\rho^{Ch.5}_{thickness} = 0.676$. In this study, the same correlation is

estimated to be $\rho_{thickness} = 0.643$ ($p = 0.021$) when including data from all 13 patients. It can be seen in the figure that, as initially observed in Chapter 5, the agreement of the data to a spherical geometric model is poor. For patients on treatment, the geometric mean change in disease volume between scan dates was +6.8%, and the corresponding change in summed linear thickness was +2.9%. For patients on observation, the geometric mean change in disease volume between scan dates was +3.9%, and the corresponding change in summed linear thickness was +0.9%.

6.5.2 DCE-CT Parameter Changes

To quantify the mean changes in DCE-CT parameter values between the first and second scan, the DCE-CT changes were first averaged over a given patient's regions of interest. Next, the unique patients in a given classification (i.e., treatment or observation) were averaged together. The results are shown in Table 6.4. Differences in mean DCE-CT parameter changes between the treatment and observation group were quantified using Student's t-test. Perhaps the most noteworthy DCE-CT parameter in the table is perfusion; for patients undergoing treatment, perfusion declines by an average of 12.6% between scans, while perfusion increases by an average of 29.4% for patients on observation. This finding is in line with the expectation that biologically active therapy reduces the proliferative cell burden, and therefore the blood flow to the region would also be reduced. None of the p-values in the table are significant, in large part due to the small number of patients in each classification group.

The changes in DCE-CT parameter values are plotted against the changes in tumor bulk measurements between the two scan dates in Figures 6.13 through 6.17. Table 6.5 shows the calculated rank correlation statistics (with respective p-values for a null hypothesis of $\rho = 0$) comparing changes in DCE-CT parameter values with changes in tumor bulk as measured with either disease volume segmentations or modified RECIST measurements. For the figures mentioned above, note that there are multiple changes in each DCE-CT parameter per patient (one for each ROI) but only one change in tumor bulk per patient. This multiplicity is due to the per-ROI nature of DCE-CT

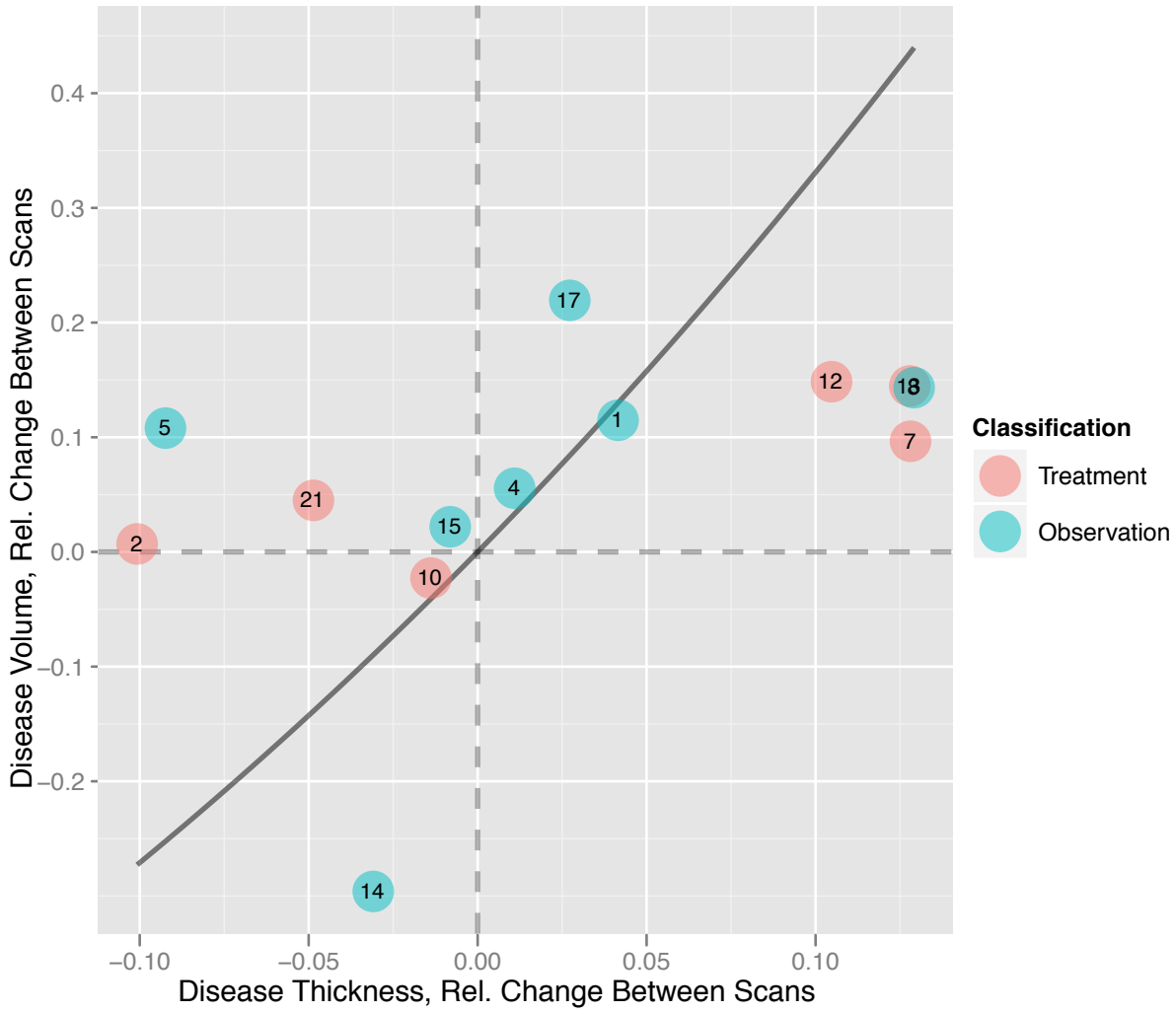


Figure 6.12: Correlation in relative change between scan dates for disease volume measurements and summed linear thickness measurements. The relationship expected from a spherical geometric model (see Chapter 5) is indicated with the solid gray line. Each point is labeled with the corresponding patient number.

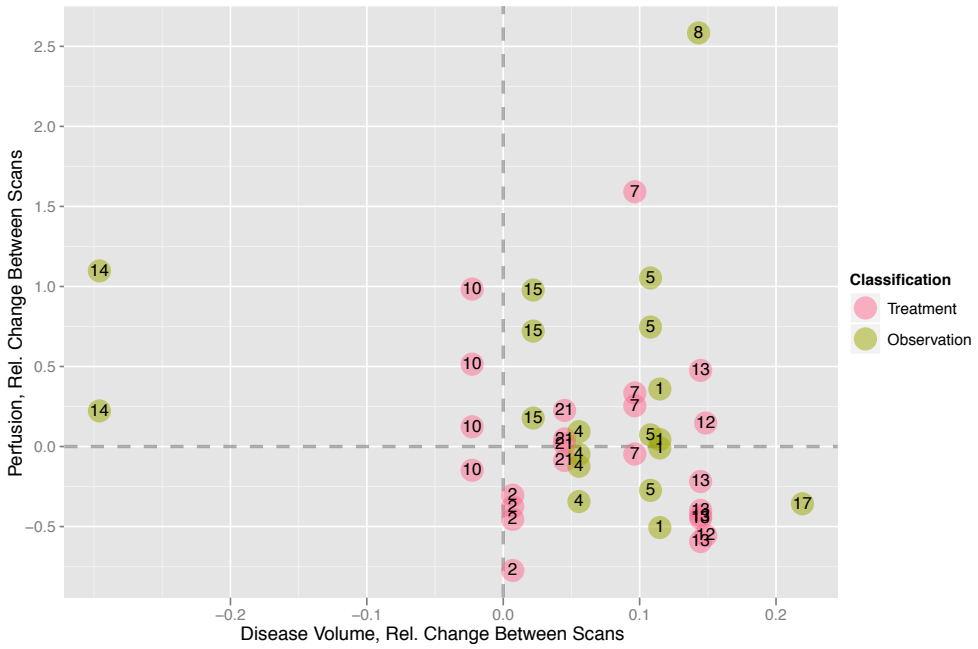
DCE-CT Parameter	Δ, All Patients	Δ, Treated Patients	Δ, Observation Patients	p-value, difference between groups
Perfusion	+7.9%	-12.6%	+29.4%	0.20
Peak Enhancement	-8.6%	-7.6%	-9.5%	0.99
Blood Volume	-7.2%	-12.8%	-2.2%	0.50
Time To Peak	-15.2%	-9.9%	-19.4%	0.77

Table 6.4: Average changes (Δ) in DCE-CT parameters from first scan to second scan. The p-values are calculated using a Student's t-test. The mean transit time parameter is not calculable for all ROI average uptake curves, resulting in insufficient data to calculate average changes.

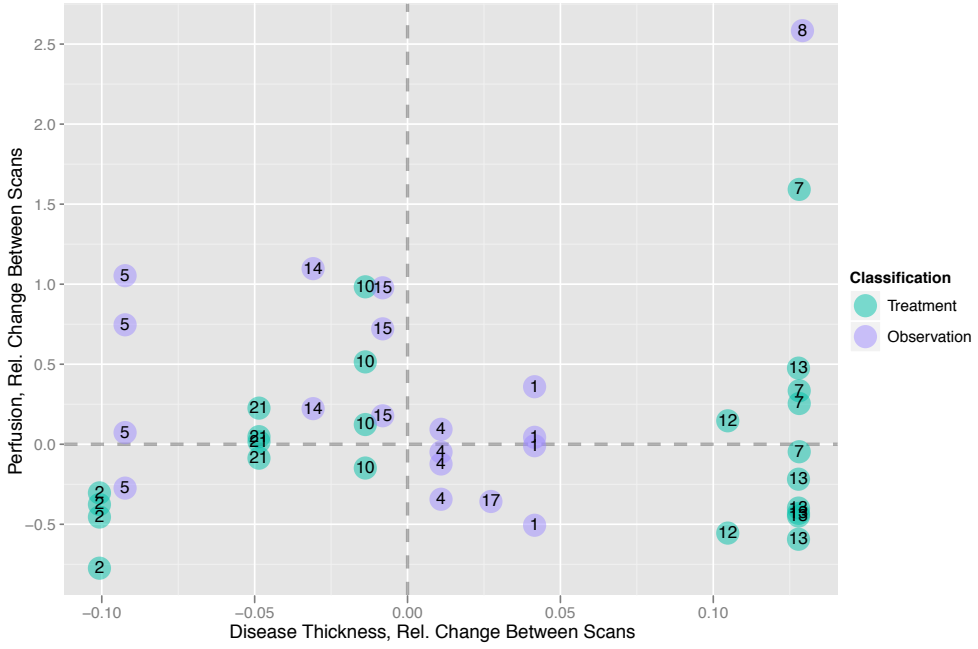
parameter calculation and the per-patient nature of the tumor bulk measurement. Rank correlation values and corresponding p-values were calculated using the per-patient average change in each DCE-CT parameter, as in Table 6.4. There are no obvious patterns in these data, as summarized with the low correlation values found in Table 6.5, though as mentioned previously, there was no *a priori* expectation of high correlation.

6.6 Discussion

The current study is part of an ongoing pilot study to explore the feasibility of dynamic contrast-enhanced computed tomography and the information contained in the DCE-CT parameter values for patients with malignant pleural mesothelioma. In the current work, we have established that DCE-CT integrated with a standard chest CT scan is a feasible imaging tool for gathering anatomic and physiologic information in MPM patients and that first-pass hemodynamic parameters can be calculated from the resulting dynamic image data. To date, 13 patients have received both the first and second planned DCE-CT scans, while an additional three patients await their second scan.

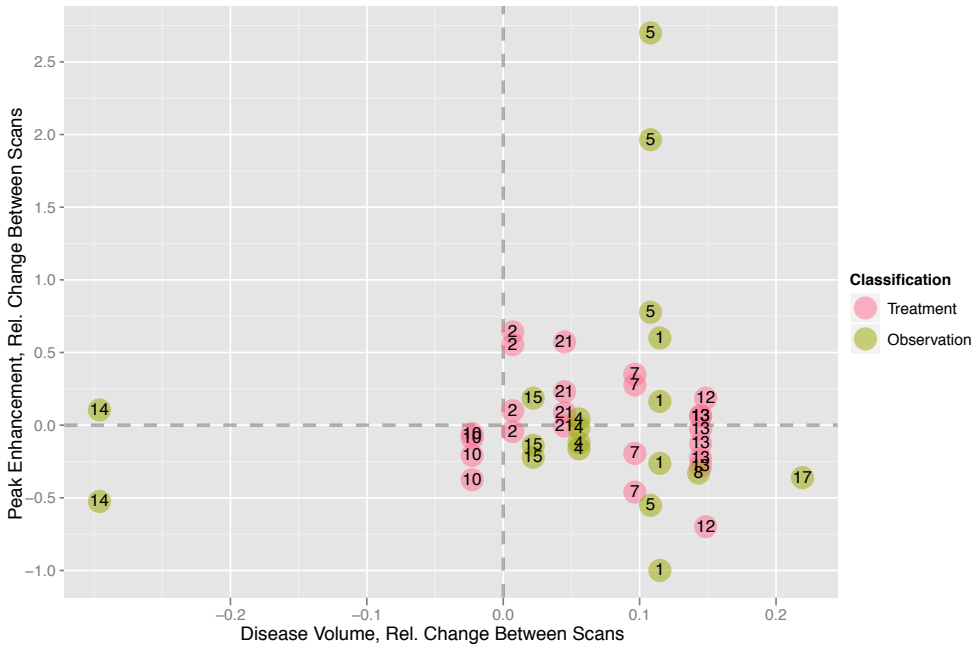


(a)

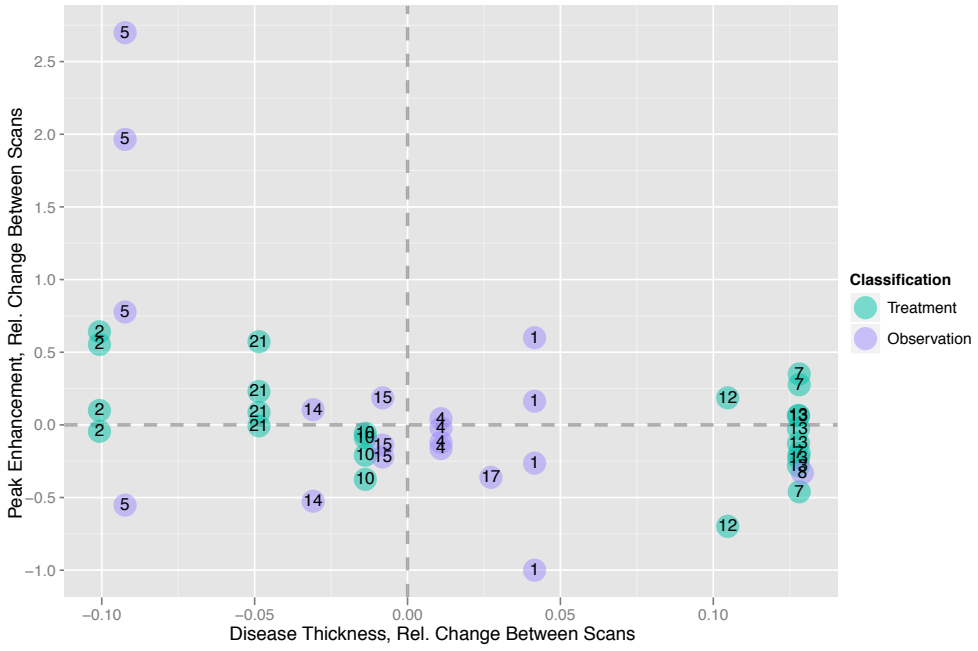


(b)

Figure 6.13: Changes in DCE-CT perfusion values versus changes in tumor bulk as measured with either (6.13a) disease volume or (6.13b) modified RECIST measurements. Each point is labeled with the corresponding patient number.

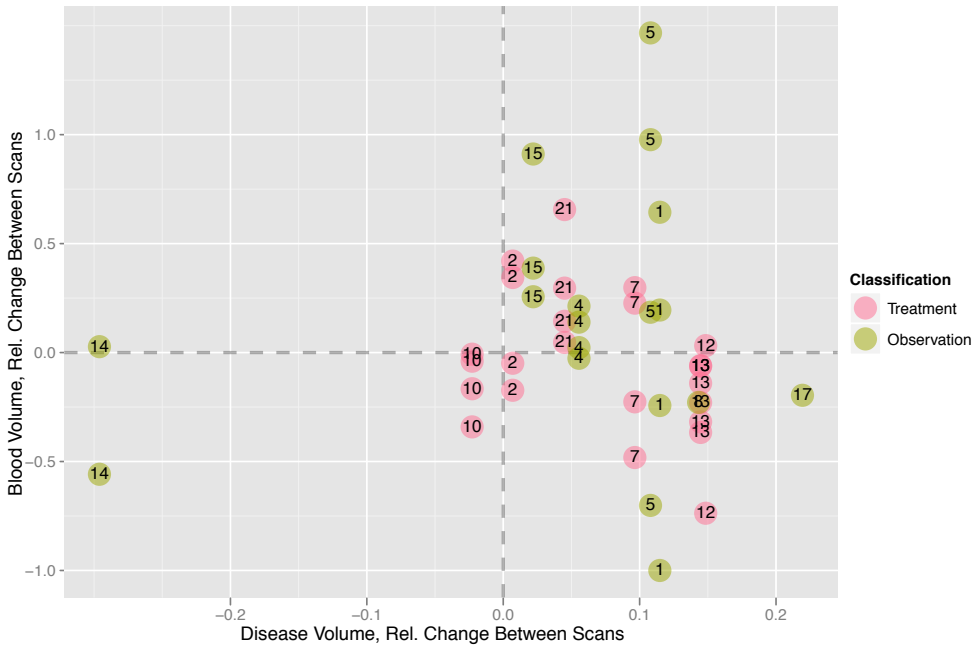


(a)

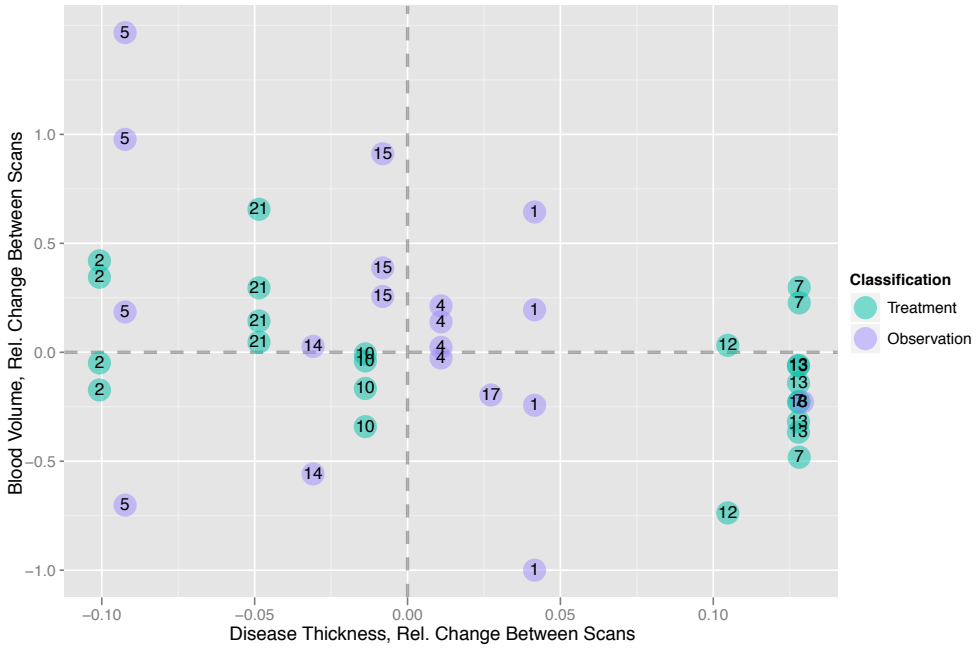


(b)

Figure 6.14: Changes in DCE-CT peak enhancement values versus changes in tumor bulk as measured with either (6.14a) disease volume or (6.14b) modified RECIST measurements. Each point is labeled with the corresponding patient number.

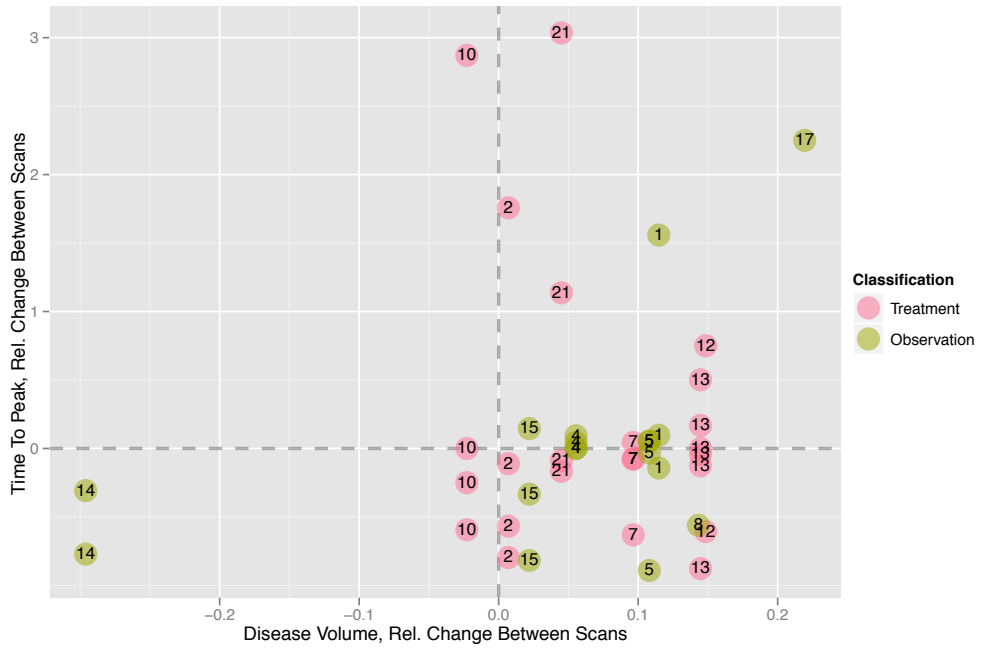


(a)

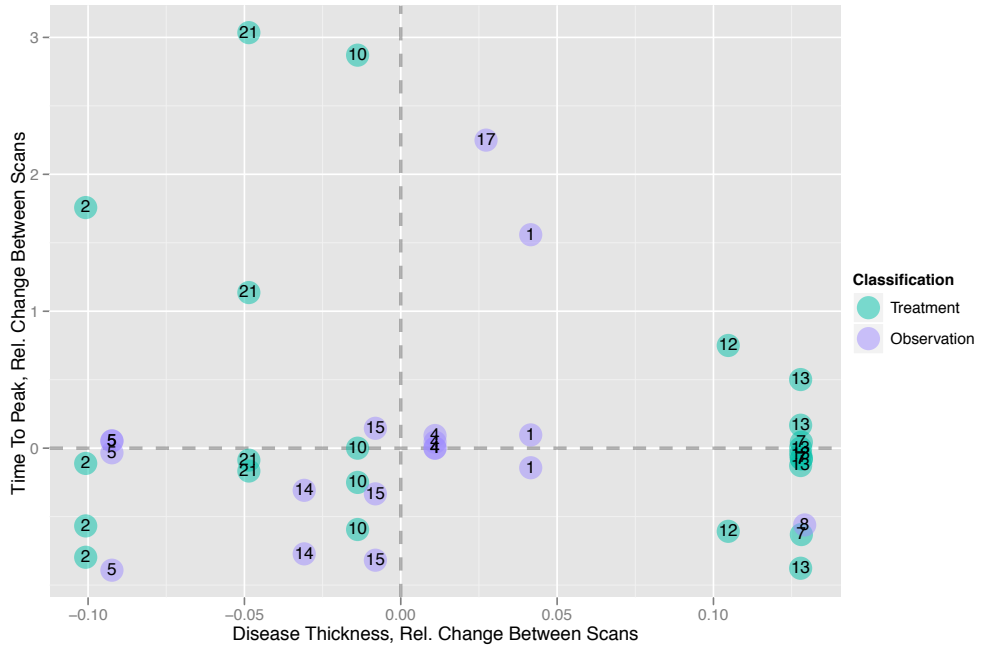


(b)

Figure 6.15: Changes in DCE-CT blood volume values versus changes in tumor bulk as measured with either (6.15a) disease volume or (6.15b) modified RECIST measurements. Each point is labeled with the corresponding patient number.

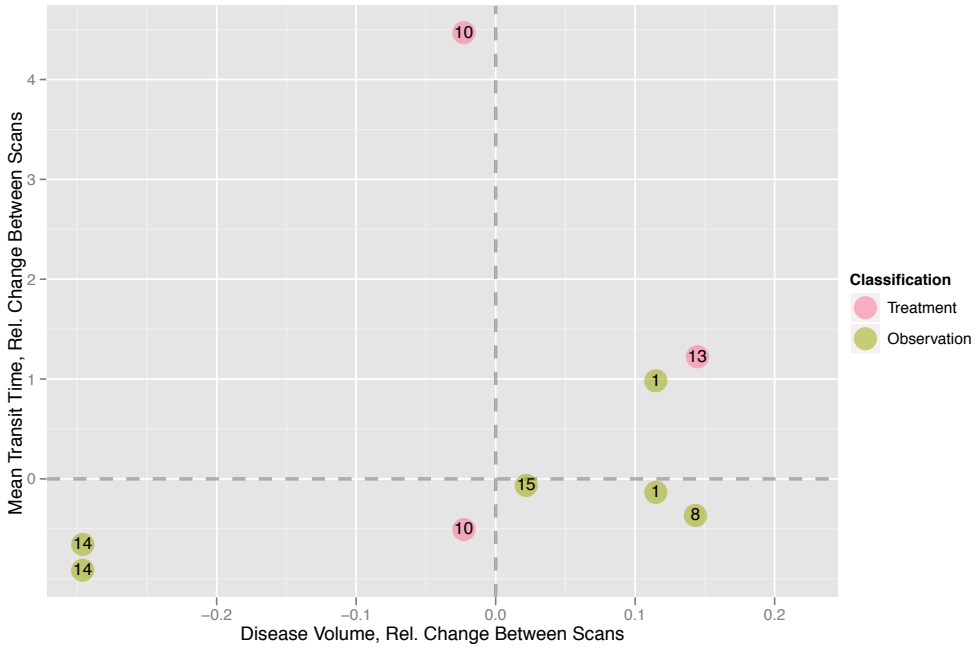


(a)

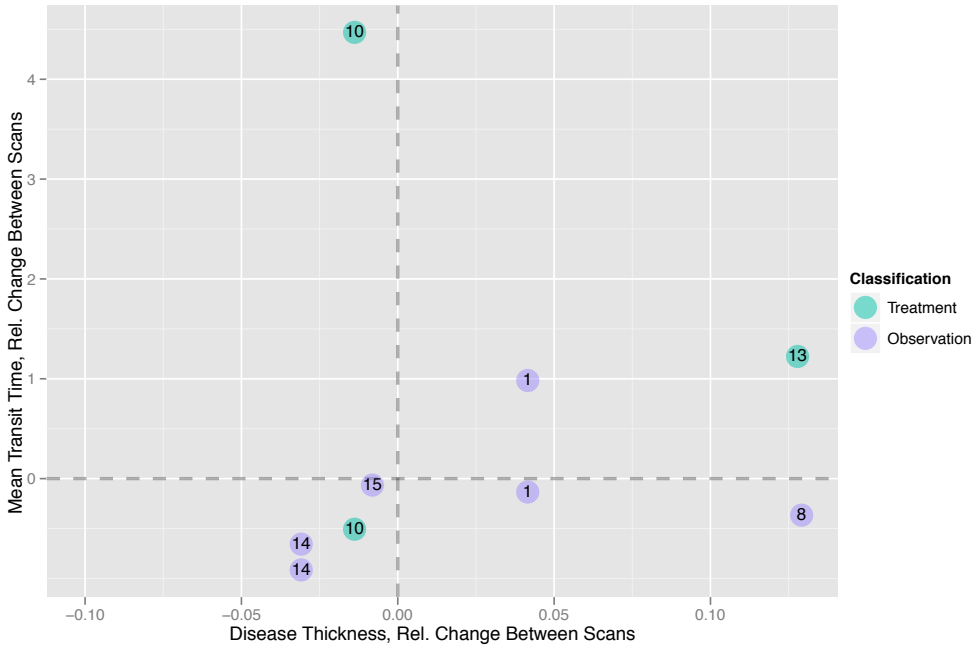


(b)

Figure 6.16: Changes in DCE-CT time to peak values versus changes in tumor bulk as measured with either (6.16a) disease volume or (6.16b) modified RECIST measurements. Each point is labeled with the corresponding patient number.



(a)



(b)

Figure 6.17: Changes in DCE-CT mean transit time values versus changes in tumor bulk as measured with either (6.17a) disease volume or (6.17b) modified RECIST measurements. Each point is labeled with the corresponding patient number. The mean transit time parameter is not calculable for all ROI average uptake curves.

DCE-CT Parameter	Tumor Measurement Technique	ρ, All Patients	ρ, Treated Patients	ρ, Observation Patients
Perfusion	Volume	-0.385 ($p = 0.20$)	-0.143 ($p = 0.80$)	-0.357 ($p = 0.44$)
	RECIST	0.148 ($p = 0.63$)	0.486 ($p = 0.36$)	-0.036 ($p = 0.96$)
Peak Enhancement	Volume	-0.363 ($p = 0.22$)	-0.429 ($p = 0.42$)	-0.321 ($p = 0.50$)
	RECIST	-0.566 ($p = 0.05$)	-0.486 ($p = 0.36$)	-0.429 ($p = 0.35$)
Blood Volume	Volume	-0.352 ($p = 0.24$)	-0.600 ($p = 0.24$)	-0.107 ($p = 0.84$)
	RECIST	-0.544 ($p = 0.06$)	-0.543 ($p = 0.30$)	-0.286 ($p = 0.56$)
Time To Peak	Volume	0.308 ($p = 0.31$)	-0.143 ($p = 0.80$)	0.607 ($p = 0.17$)
	RECIST	0.033 ($p = 0.92$)	-0.086 ($p = 0.92$)	0.286 ($p = 0.56$)

Table 6.5: Rank correlation statistics and respective p-values comparing changes in DCE-CT parameter values with changes in tumor bulk as measured with either disease volume segmentations or modified RECIST measurements. All p-values are calculated for a null hypothesis of $\rho = 0$. The mean transit time parameter is not calculable for all ROI average uptake curves, resulting in insufficient data to calculate rank correlation values.

Patient accrual is planned to continue for an additional four patients, leading to a total of 20 patients (10 each on treatment and on observation).

The imaging protocol in this study was initially developed by studying other previously implemented DCE-CT protocols for thoracic malignancies, namely lung cancer. The imaging parameters (120 kVp and 100 mAs) provided sufficient contrast and low enough noise for 3-mm axial reconstructions to be easily visually interpretable. Later, after a quantitative study of imaging dose and image quality metrics, it was decided that 100 kVp and 100 mAs provided a better balance between image quality and patient dose. Imaging parameters of 120 kVp and 80 mAs were "a close second," with similar imaging dose values between the two protocols, but the enhanced iodine attenuation of the 100 kVp spectrum leads to a higher signal from the contrast uptake in vascular tissues. While noise is somewhat increased by using a lower tube voltage setting, the increase in noise does not preclude visual interpretation or image registration (while no patient in this written document has received the "updated" imaging protocol, two of the patients awaiting their second DCE-CT scan have been scanned with 100 kVp/100 mAs). Furthermore, since DCE-CT analysis is predominantly performed on average uptake curves, the process of spatially averaging HU values for each DCE-CT snapshot in a masked ROI region compensates for any image noise in the individual DCE-CT snapshots (the smooth appearance of the average uptake curves in Figure 6.5 are an example of this effect).

The software toolkit and workflow developed specifically for this study allow for analysis of volumetric regions of interest of arbitrary morphology in a spatially co-registered coordinate system (chosen for this study as the 20th DCE-CT snapshot). The application of non-rigid image registration for DCE-CT is not novel, since its use has been described elsewhere [129–131]. In this study, registration to a common "central" snapshot was performed, rather than "daisy-chaining" the registrations together (e.g., $1 \rightarrow 20$ directly rather than a composite deformation vector field generated from $1 \rightarrow 2$, $2 \rightarrow 3$,..., $19 \rightarrow 20$). While this technique was developed independently for this study after exploration of our own DCE-CT image data, the technique is in fact recommended

in recent literature [129]. The non-rigid registration technique used in this study (the symmetric normalization algorithm in ANTs) is relatively new and has been seemingly used in only one other dynamic imaging study [142]. Its use in this study was motivated by the qualitative improvements in deformed image appearance and reduction in deformation artifacts when compared with the demons algorithm implementation. A quantitative comparison of residual error in matched anatomical landmarks would conceivably show similar performance in spatial matching between demons and ANTs, since both techniques achieve good spatial matching across DCE-CT snapshots. It should be noted that the onus is still on the observer to identify any gross mis-registrations and draw regions of interest that avoid these areas.

When calculating changes in DCE-CT parameters between successive scans, it is important to match the definitions of spatial regions of interest. Initially, we attempted to register all the DCE-CT snapshots from the second DCE-CT scan to the 20th snapshot of the first DCE-CT scan. Had this approach been successful, the ROI contours from the first DCE-CT scan could be used on the second DCE-CT scan without too much editing. However, due to potentially gross positioning differences between DCE-CT scans, this attempt was often not successful (or if partially successful, the useful axial range might potentially be reduced from 54 mm because of changes in patient pitch in the scanner bore). Instead, the strategy was adopted to independently contour matched ROIs on the second DCE-CT scan for each patient with reference to the first DCE-CT scan.

There were no obvious correlations between changes in tumor bulk measurements and changes in DCE-CT parameter values in Table 6.5, and no individual correlation was significantly different from zero. While the rank correlation tests could have been performed using per-ROI changes in DCE-CT parameter values rather than the average per-patient changes, "response" is inherently a per-patient phenomenon, and changes in tumor measurements are reported on a per-patient basis. In Table 6.4, the DCE-CT parameter perfusion showed an interesting separation between the treatment and observation patient cohorts. For patients on treatment, perfusion decreased on average between the two DCE-CT scans, while for patients on observation, perfusion increased on aver-

age. It will be interesting to investigate this parameter again when the full 20 patients have been scanned. For the two mesothelioma patients on the anti-angiogenic therapy AZD2171 in the study by Meijerink *et al*. [80], there was a reduction in perfusion between DCE-CT scan dates matched with a decrease in tumor bulk as measured using the modified RECIST technique (one patient experienced a perfusion reduction of 19%, while the other experienced a reduction of 47%).

It should be noted that the patient cohort for this study is small. Trends within patient subsets (six patients on treatment, seven patients on observation) are tenuous, and the final analysis after 20 patients have been accrued will only be marginally better powered from a statistical perspective. The goal of this work was always intended to be "motivational" in the pilot-study sense of the word. Based on these data, a future study may potentially be designed to further investigate DCE-CT parameters for patients all undergoing the same treatment (even with six treatment patients in this study, there are four unique treatment regimens!). Since it has been shown that the complete calculation of mean transit time would likely require more imaging at the "distal end" of the current DCE-CT protocol, the revised imaging protocol could either cut the second group of DCE-CT snapshots entirely or increase the delay until their acquisition.

As survival times become available for the patients in this study, it will also be interesting to investigate correlations among DCE-CT parameters and survival. In the DCE-MRI study by Giesel *et al*. [76], the kinetic parameter k_{ep} was significantly associated with patient survival. While the parameter k_{ep} often used in DCE-MRI literature has no direct analogue in first-pass DCE-CT, it is still a normalized contrast uptake rate and is therefore most similar to perfusion [143]. It will be interesting to investigate DCE-CT perfusion as a prognostic parameter for MPM as more patient data become available.

CHAPTER 7

CONCLUSION AND FUTURE DIRECTIONS

"What is a scientist after all? It is a curious man looking through a keyhole, the keyhole of nature, trying to know what's going on." – Jacques Yves Cousteau

The work in this dissertation has been focused on the redefinition of radiologic response for patients with malignant pleural mesothelioma (MPM). The current standard for medical image-based tumor response assessment for patients with MPM is the modified Response Evaluation Criteria In Solid Tumors (RECIST) measurement technique [47] with changes classified according to the standard RECIST response classification criteria [43]. In the modified RECIST protocol, two tumor thickness (short-axis) measurements are acquired from each of three axial sections; the sum of the six linear measurements forms the tumor burden measurement to be tracked over time. Using the standard RECIST classification criteria, if the summed linear measurement increases by more than 20% from the minimum of previous measurements (nadir), the patient is classified as having progressive disease (PD), and if the measurement decreases by 30% or more from the baseline measurement, the patient is classified as partially responsive (PR). Stable disease (SD) is defined for measurement changes lying between the previously mentioned classification groups. While the modified RECIST measurement technique was developed specifically for the unique morphology and growth patterns of mesothelioma, the standard RECIST classification criteria are used across a wide range of diseases and are certainly not specific to any one disease. In fact, the standard RECIST classification criteria are extrapolated from an assumption of spherical geometry and the amount of change physicians believed they could reliably identify from palpation alone in previous breast cancer cohort studies [44].

The individual projects that form the chapters of this dissertation extend the current response assessment paradigm by investigating alternate response assessment strategies in one, two, three, and four dimensions. Tumor burden measurements were extended from one to three dimensions, and tumor response was extended from discrete to continuous quantification of change in tumor

burden. Normal tissue structures were investigated as potential surrogates of tumor response. Finally, temporal and spatial measurements were combined in an investigation of dynamic imaging for patients with MPM.

In Chapter 2, the performance of the current response classification criteria was quantified and new discrete response classification criteria were optimized for patients with MPM using the current linear (1D) measurement technique. Seventy-eight patients were included from of a patient database consisting of 97 treated patients from Sir Charles Gairdner Hospital in Perth, Western Australia. Using Harrell's C statistic, the performance of the current classification criteria (i.e., -30%/+20%) and each patient's best response during treatment was $C_{std} = 0.778$. Despite the original arbitrary provenance of these classification cut points, they perform adequately and are within the range of "clinical utility." However, the most useful (and relevant) response classification criteria for any given disease are those developed to maximize the association between the assessment metric and meaningful outcomes such as survival. In this study, the optimal association between tumor response and patient survival was obtained using classification criteria of -64% and +50% for PR and PD, respectively. Using the same Harrell's C statistic, the performance of the new optimized criteria (-64%/+50%) and each patient's best response was $C_{opt} = 0.855$. These optimized classification criteria appear better suited to the specific morphology and growth pattern of mesothelioma and may prove useful in the assessment of clinical trials and routine patient care. Furthermore, these optimized criteria allow for a marginal improvement (though not significant; $p = 0.12$ in a cross-validation) in response assessment performance while maintaining the simple linear thickness tumor burden measurement technique.

Because of the limited number of patients in the database, these optimized classification criteria need to undergo a full validation on an independent patient cohort. The new criteria are similar to theoretical linear measurement classification criteria derived for non-spherical geometries by Oxnard *et al*. [51], but because the classification criteria would impact clinical trial eligibility and triggers to cease clinical trial treatment, it is important that the new classification criteria be vali-

dated before utilization in a clinical setting. Furthermore, the optimization process may need to be repeated for patients undergoing therapeutic regimens beyond the cytotoxic therapy of the patients in this study.

In Chapter 3, manual area (2D) contours were investigated as a potential means of response assessment in MPM patients. In place of summed tumor thickness measurements, summed area measurements from three axial sections are potentially better able to capture morphological changes in tumor burden while avoiding the time cost associated with manual segmentation of complete disease volume. The goal of the study was to evaluate tumor area measurements on CT scans as a more complete and, potentially, less variable metric for MPM response assessment than linear thickness measurements acquired according to the modified RECIST protocol. Inter-observer variability was initially quantified for baseline scans (i.e., at a single time-point), where the relative variability between observers' summed area measurements spanned -71% – 240% (95% confidence interval). Observers were also tasked with drawing follow-up contours on a CT scan from a later time-point for each patient, where the baseline scan and contours were visible to the observers for each patient during the follow-up contouring task. The baseline contours implicitly constrained the follow-up measurement process, and the relative variability between observers was reduced to -41% – 69% for the follow-up summed area measurements (95% confidence interval).

The measured inter-observer variability in both baseline and follow-up summed area measurements exceeds two-dimensional response classification criteria and are much wider than corresponding inter-observer variabilities previously published for summed linear thickness measurements [52, 53]. Therefore, while the number of patients (31) and observers (five for baseline, three for follow-up) in this study are relatively low, the primary implication of this study is that summed area measurements are too variable for reliable use as a response assessment technique in patients with MPM. Beyond the variability in the data, the collection of area contours from three axial sections per patient is considerably more time-consuming than the linear measurement technique currently utilized clinically. The time burden per patient is similar to the time required to complete

a semi-automated disease volume segmentation, and therefore it is difficult to foresee a situation in which summed area measurements would be used in place of semi-automated segmentations of actual disease volume for response assessment.

In Chapter 4, a comprehensive model to predict MPM patient survival was built using time-changing measurements of image-based disease volume in conjunction with clinical covariates. While prognostic models have been published for patients with MPM previously, they have either included clinical covariates or measurements of disease volume, never both. Furthermore, the only model incorporating *changes* in disease volume used discrete change (i.e., disease growth or disease decline) assessed only once (i.e., between baseline and first follow-up scans) [58]. In the present study, pleural disease was segmented using a semi-automated technique with modifications from previously published literature. Changes in pleural disease volume were treated continuously and were quantified using the "specific growth rate" (SGR) evaluated over time. Patient survival was modeled using a time-variable implementation of the Cox Proportional Hazards model, and survival model performance was assessed using Heagerty's C^{τ} statistic, an extension of receiver operating characteristic (ROC) analysis to time-dependent survival models.

Using the semi-automated segmentation method customized for this study, final segmentations of pleural disease volume were obtained with approximately 20 minutes of manual intervention per case, compared with around 45 minutes of manual intervention per case using the previously published "automated" segmentation technique. In the cohort of 81 patients (281 CT scans), disease volume SGR was significantly associated with patient survival in a univariate model ($p = 0.0003$), as were eight clinical covariates in univariate analysis. The final multivariate model for patient survival used disease volume SGR ($p = 0.00045$) in conjunction with disease histology ($p = 0.0029$), dyspnea ($p = 0.0020$), and Eastern Cooperative Oncology Group (ECOG) performance status ($p = 0.029$), where again disease volume was allowed to vary over time as disease volume measurements were accumulated from a series of follow-up CT scans. The performance of the survival model, when trained and tested on the same full patient cohort, was estimated to be

$C^{\tau} = 0.69$, whereas the performance estimate was reduced to $C^{\tau} = 0.66$ when model performance was cross-validated using either a leave-one-out approach or repeated random sub-sampling.

The multivariate survival model in this study is the only model developed to date that includes both clinical covariates and time-variable measurements of disease volume for patients with MPM. While changes in disease volume were previously shown to be associated with patient survival when discretized between growth and decline, these changes were only assessed at a single time-point, and subsequent updates from follow-up scans during the entire course of a patient's treatment were necessarily disregarded. When modeled according to the methods of Liu *et al*. [58], patient survival only achieved a performance of $C^{\tau} = 0.521$. However, any arbitrary disease-volume trajectory can be modeled using the techniques in this study, and model performance is improved accordingly.

Future work will focus on expanding the patient cohort and validating the prognostic model in independent databases. Efforts are currently underway to use semi-automated disease volume measurements for response assessment in a clinical trial for patients with MPM. While the patients in the current study all received cytotoxic chemotherapy, the number of cycles per patient was prescribed clinically, and drug selection was also based on clinical standard of care. It will be interesting to see whether the model from this study will be applicable to a new treatment regimen, or whether treatment-specific models will need to be developed (e.g., for a targeted anti-angiogenic treatment). Furthermore, future work could investigate incorporating disease volume measurements from the pre-treatment period. Though not investigated in the current study due to time constraints, changes in disease SGR between the pre-treatment period and the during-treatment period could be indicative of patient response.

In Chapter 5, automatically segmented *lung* volumes were investigated as an alternate response assessment measurement technique. Because of the anatomy associated with MPM, changing tumor burden affects more than just the volume of tumor. The hemithoracic space is fairly fixed by the ribcage and mediastinal anatomy. Therefore, when the disease volume increases, it is rea-

sonable to believe that the aerated lung volume should decrease correspondingly. This correlation was observed for data in an applicable cohort of 61 patients (a subset of the cohort from Chapter 4), where the rank correlation between changes in disease volume and aerated lung volume was $\rho = -0.69$. Furthermore, while increasing disease volume was shown to be significantly associated with poor patient prognosis in the previous chapter, this study revealed decreases in lung volume to be significantly associated with poor patient prognosis (as expected). When modeled as time-variable SGR, disease volume and lung volume were nearly identical in terms of their statistical prognostic performance (as assessed using C^{τ} in multivariate survival models).

Linear measurements were also compared with disease volumes in Chapter 5, where again a high rank correlation between changes in linear thickness and disease volume was observed ($\rho = 0.68$). Summed linear thickness measurements were significantly associated with patient survival ($p = 0.00053$) when modeled as time-variable SGR in conjunction with the previously-mentioned clinical covariates. Prognostic performance was on average better in multivariate survival models using summed linear thickness measurements than with models using disease volume measurements, though the difference was not statistically significant in a repeated random sub-sample analysis.

The broad strokes of Chapter 5 are quite similar to those of Chapter 4: relevant measurements (whether of the disease burden directly or of surrogate normal structures), treated continuously and modeled as time-variable specific growth rates, are significantly associated with patient survival in multivariate models. However, while the statistical performance of the three survival models are very similar, there are certainly practical differences between the three measurement techniques. Summed linear thickness measurements are obtained manually and require perhaps five minutes of manual intervention per case, whereas disease volumes are obtained semi-automatically, requiring around 20 minutes per case of manual intervention. Lung volumes are obtained completely automatically with no user invention required and therefore are perhaps the most desirable of the three measurement techniques. However, lung volumes were only investigated as a response assessment

measurement technique in patients with unilateral disease since, to be tracked over time, ipsilateral lung volumes need to be normalized by contralateral healthy lung volumes; lung volumes are not applicable for response assessment in patients with bilateral disease (around 10% of patients). While survival models using summed linear thickness measurements performed better (on average) than corresponding models using disease volumes, Frauenfelder *et al*. [57] reported considerably lower inter-observer variability in disease volume measurements than in summed linear thickness measurements. The result may have been in part due to their elementary volumetric measurement technique and arbitrary response criteria with a very wide definition of stable disease, but taken at face value, the reduced inter-observer variability of volume measurements should temper the apparent advantage of linear thickness measurements.

While disease volumes are logically better able to capture changes in overall tumor bulk, it is possible that changes in tumor thickness may be more directly associated with patient prognosis than overall disease volume, leading to the observed performance increase for the survival model using summed linear thickness measurements. One possible explanation is that human observers are able to place their baseline tumor thickness measurements in locations that are in some sense "important" for response assessment; volume measurements capture changes over the total extent of disease, while summed tumor thickness measurements only capture change in the discrete (up to six, by modified RECIST) locations where baseline measurements were placed. This hypothesis could be tested by using the available volumetric segmentations to derive automated random linear thickness measurements at baseline. Follow-up measurements could be made by picking matched locations using deformable registration and again using the volumetric segmentation to identify the local "tumor thickness." The prognostic performance of these effective linear measurements could be compared with the performance of the existing manual measurements, indicating to what extent manual measurement placement was better able to predict patient survival than "random" linear thickness measurements.

Whether changes in tumor burden are assessed using summed linear thickness measurements,

disease volume measurements, or normalized lung volume measurements, many recommend that tumor response should be assessed using continuous (i.e., not discrete) measurement changes [44, 45, 106, 144]. Therefore, while a discretized model for tumor response was optimized specifically for patients with MPM in Chapter 2, tumor response should ultimately rely on continuous measurement changes. Rather than defining study eligibility and exclusion criteria in terms of discretized response classification, continuous SGR measurements may be better able to identify when patients are in need of therapeutic intervention or are no longer gaining benefit from an administered therapy. This dissertation provides the first evidence that continuous changes in all three measurement techniques are significantly associated with patient survival.

Finally, Chapter 6 reported the results of a pilot study on dynamic contrast-enhanced (DCE) CT for patients with MPM. Of the 13 patients with two completed DCE-CT scans included in this initial analysis, six were on treatment and seven were on observation. The final accrual goal is 10 patients in each classification group with two DCE-CT scans each. A DCE-CT imaging protocol was developed to dovetail with a clinically indicated standard chest CT scan and utilize the same bolus of iodinated contrast media. A deformable registration software tool was implemented to co-register the individual DCE-CT snapshots in a single scan to a common reference frame, allowing contrast uptake to be tracked on a voxel-by-voxel level. Volumetric foci of disease were contoured separately, and various hemodynamic parameters were calculated for each region of interest. DCE-CT parameter values are reported on a per-scan basis, as well as changes in the DCE-CT parameters for each patient group.

There were no obvious correlations between changes in tumor bulk measurements and changes in DCE-CT parameter values. While measurements of tumor burden are often used as tools for response assessment, changes in tumor burden may not be directly related to changes in the hemodynamic properties of the disease, and there was no *a priori* expectation of high correlation between changes in tumor burden and changes in DCE-CT parameters. The DCE-CT parameter perfusion showed an interesting separation between the treatment and observation patient cohorts.

For patients on treatment, perfusion decreased by 13% on average between the two DCE-CT scans, while for patients on observation, perfusion increased by 29% on average. It will be interesting to investigate this parameter again as a potential indicator of response when the full 20 patients have been scanned. Another parameter of interest was the time until peak enhancement, which averaged around 86 seconds after initiating contrast injection. While the typical delay between contrast injection and standard chest CT scan acquisition is around 60–70 seconds, clinicians may be able to improve single-snapshot MPM disease enhancement by extending the duration of the delay period.

It should be noted that the patient cohort in Chapter 6 is small. Trends within patient subsets (six patients on treatment, seven patients on observation) are tenuous, and the final analysis after 20 patients have been accrued will only be marginally better powered from a statistical perspective. The goal of this work was always intended to be "motivational" in the pilot-study sense of the word. Based on these data, a future study may potentially be designed to further investigate DCE-CT parameters for patients all undergoing the same treatment (even with six treatment patients in this study, there are four unique treatment regimens represented).

One feasible and interesting direction of future work motivated by this particular study would be the investigation of correlations between DCE-CT data and fluorodeoxyglucose positron emission tomography (FDG-PET) data. Some of the patients imaged in this study with DCE-CT have also been imaged at our institution with FDG-PET, and DCE-CT parameters could be correlated with FDG-PET imaging parameters such as the standardized uptake value (SUV). FDG-PET imaging parameters have been shown to be significantly associated with patient survival in multiple studies [25, 28, 35, 37, 38], and correlations between DCE-CT and FDG-PET parameters may provide insight into which DCE-CT parameters are potentially prognostic before survival data becomes available on the current patient cohort.

As a coherent progression, this work has extended the definition of radiologic response for patients with malignant pleural mesothelioma. New classification criteria were obtained for one-dimensional tumor measurements and improved the classification performance of the discrete re-

sponse classification model. Two-dimensional area measurements were investigated as a potential tumor measurement technique and were found to be too variable for reliable response assessment. Three-dimensional measurements of disease and lung volume were found to be associated with patient survival, and survival models using linear measurements were in fact found to out-perform the models using three-dimensional measurements. Finally, dynamic CT imaging was investigated as a potential four-dimensional imaging technique for patients with mesothelioma and DCE-CT parameters were calculated from the dynamic imaging data. My sincere hope is that the benefit to clinicians and patients will match the enjoyment obtained in researching these projects.

REFERENCES

[1] B. W. S. Robinson and R. A. Lake. Advances in malignant mesothelioma. *N Engl J Med*, 353:1591–1603, 2005.

[2] H. Pass, N. J. Vogelzang, S. Hahn, and M. Carbone. Malignant pleural mesothelioma. *Curr Probl Cancer*, 28(3):93–174, 2004.

[3] H. Weill, J. M. Hughes, and A. M. Churg. Changing trends in US mesothelioma incidence. *Occup Environ Med*, 61:438–441, 2004.

[4] V. L. Roggli, A. Sharma, K. J. Butnor, T. Sporn, and R. T. Vollmer. Malignant mesothelioma and occupational exposure to asbestos: A clinicopathological correlation of 1445 cases. *Ultrastruct Pathol*, 26(2):55–65, 2002.

[5] R. W. Carlson, D. C. Allred, B. O. Anderson, H. J. Burstein, W. B. Carter, S. B. Edge, J. K. Erban, W. B. Farrar, L. J. Goldstein, W. J. Gradishar, D. F. Hayes, C. A. Hudis, M. Jahanzeb, K. Kiel, B.-M. Ljung, P. K. Marcom, I. A. Mayer, B. McCormick, L. M. Nabell, L. J. Pierce, E. C. Reed, M. L. Smith, G. Somlo, R. L. Theriault, N. S. Topham, J. H. Ward, E. P. Winer, A. C. Wolff, and NCCN Breast Cancer Clinical Practice Guidelines Panel. Breast cancer: Clinical practice guidelines in oncology. *J Natl Compr Canc Netw*, 7(2):122–192, 2009.

[6] S. Halabi, E. J. Small, P. W. Kantoff, M. W. Kattan, E. B. Kaplan, N. A. Dawson, E. G. Levine, B. A. Blumenstein, and N. J. Vogelzang. Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. *J Clin Oncol*, 21(7):1232–1237, 2003.

[7] J. Peto, A. Decarli, C. La Vecchia, F. Levi, and E. Negri. The European mesothelioma epidemic. *Br J Cancer*, 79:666–672, 1999.

[8] B. P. Lanphear and C. R. Buncher. Latent period for malignant mesothelioma of occupational origin. *J Occup Med*, 34(7):718–721, 1992.

[9] A. S. Tsao, I. Wistuba, J. A. Roth, and H. L. Kindler. Malignant pleural mesothelioma. *J Clin Oncol*, 27:2081–2090, 2009.

[10] T. Treasure, L. Lang-Lazdunski, D. Waller, J. M. Bliss, C. Tan, J. Entwisle, M. Snee, M. O'Brien, G. Thomas, S. Senan, K. O'Byrne, L. S. Kilburn, J. Spicer, D. Landau, J. Edwards, G. Coombes, L. Darlison, and J. Peto. Extra-pleural pneumonectomy versus no extra-pleural pneumonectomy for patients with malignant pleural mesothelioma: Clinical outcomes of the Mesothelioma And Radical Surgery (MARS) randomised feasibility study. *Lancet Oncol*, 12(8):763–772, 2011.

[11] V. W. Rusch, S. Piantadosi, and E. C. Holmes. The role of extrapleural pneumonectomy in malignant pleural mesothelioma: A Lung Cancer Study Group trial. *J Thorac Cardiovasc Surg*, 102(1):1–9, 1991.

[12] D. J. Sugarbaker, J. P. Garcia, W. G. Richards, D. H. Harpole, E. Healy-Baldini, M. M. DeCamp, S. J. Mentzer, M. J. Liptay, G. M. Strauss, and S. J. Swanson. Extrapleural pneumonectomy in the multimodality therapy of malignant pleural mesothelioma: Results in 120 consecutive patients. *Ann Surg*, 224(3):288–296, 1996.

[13] P. A. Jänne and E. H. Baldini. Patterns of failure following surgical resection for malignant pleural mesothelioma. *Thorac Surg Clin*, 14(4):567–573, 2004.

[14] A. Ahamad, C. W. Stevens, W. R. Smythe, A. A. Vaporciyan, R. Komaki, J. F. Kelly, Z. Liao, G. Starkschall, and K. M. Forster. Intensity-modulated radiation therapy: A novel approach to the management of malignant pleural mesothelioma. *Int J Radiat Oncol Biol Phys*, 55(3):768–775, 2003.

[15] A. M. Allen, M. Czerminska, P. A. Jänne, D. J. Sugarbaker, R. Bueno, J. R. Harris, L. Court, and E. H. Baldini. Fatal pneumonitis associated with intensity-modulated radiation therapy for mesothelioma. *Int J Radiat Oncol Biol Phys*, 65(3):640–645, 2006.

[16] W. Weder, R. A. Stahel, J. Bernhard, S. Bodis, P. Vogt, P. Ballabeni, D. Lardinois, D. Betticher, R. Schmid, R. Stupp, H. B. Ris, M. Jermann, W. Mingrone, A. D. Roth, A. Spiliopoulos, and Swiss Group for Clinical Cancer Research. Multicenter trial of neo-adjuvant chemotherapy followed by extrapleural pneumonectomy in malignant pleural mesothelioma. *Ann Oncol*, 18(7):1196–1202, 2007.

[17] N. J. Vogelzang, J. J. Rusthoven, J. Symanowski, C. Denham, E. Kaukel, P. Ruffie, U. Gatzemeier, M. Boyer, S. Emri, C. Manegold, C. Niyikiza, and P. Paoletti. Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. *J Clin Oncol*, 21(14):2636–2644, 2003.

[18] G. L. Ceresoli, B. Castagneto, P. A. Zucali, A. Favaretto, M. Mencoboni, F. Grossi, D. Cortinovis, G. Del Conte, A. Ceribelli, A. Bearz, S. Salamina, F. De Vincenzo, F. Cappuzzo, M. Marangolo, V. Torri, and A. Santoro. Pemetrexed plus carboplatin in elderly patients with malignant pleural mesothelioma: Combined analysis of two phase II trials. *Br J Cancer*, 99:51–56, 2008.

[19] S. G. Armato III. Computerized analysis of mesothelioma on CT scans. *Lung Cancer*, 49 Suppl 1:S41–4, 2005.

[20] A. Nowak, S. G. Armato III, G. L. Ceresoli, H. Yildirim, and R. J. Francis. Imaging in pleural mesothelioma: A review of imaging research presented at the 9th International Meeting of the International Mesothelioma Interest Group. *Lung Cancer*, 70(1):1–6, 2010.

[21] S. G. Armato III, J. Entwisle, M. T. Truong, A. Nowak, G. L. Ceresoli, B. Zhao, R. Misri, and H. L. Kindler. Current state and future directions of pleural mesothelioma imaging. *Lung Cancer*, 59:411–420, 2008.

[22] R. T. Heelan, V. W. Rusch, C. B. Begg, D. M. Panicek, J. F. Caravelli, and C. Eisen. Staging of malignant pleural mesothelioma: Comparison of CT and MR imaging. *Am J Roentgenol*, 172(4):1039–1047, 1999.

[23] R. R. Gill, V. H. Gerbaudo, D. J. Sugarbaker, and H. Hatabu. Current trends in radiologic management of malignant pleural mesothelioma. *Semin Thorac Cardiovasc Surg*, 21:111–120, 2009.

[24] N. Corson, W. F. Sensakovic, C. Straus, A. Starkey, and S. G. Armato III. Characterization of mesothelioma and tissues present in contrast-enhanced thoracic CT scans. *Med Phys*, 38(2):942–947, 2011.

[25] S. Basu, B. Saboury, D. A. Torigian, and A. Alavi. Current evidence base of FDG-PET/CT imaging in the clinical management of malignant pleural mesothelioma: Emerging significance of image segmentation and global disease assessment. *Mol Imaging Biol*, 13(5):801–811, 2011.

[26] R. T. Heelan. Staging and response to therapy of malignant pleural mesothelioma. *Lung Cancer*, 45 Suppl 1:S59–61, 2004.

[27] H. Otsuka, K. Terazawa, N. Morita, Y. Otomi, K. Yamashita, and H. Nishitani. Is FDG-PET/CT useful for managing malignant pleural mesothelioma? *J Med Invest*, 56:16–20, 2009.

[28] F. Bénard, D. Sterman, R. J. Smith, L. R. Kaiser, S. M. Albelda, and A. Alavi. Prognostic value of FDG PET imaging in malignant pleural mesothelioma. *J Nucl Med*, 40(8):1241–1245, 1999.

[29] V. W. Rusch. A proposed new international TNM staging system for malignant pleural mesothelioma from the International Mesothelioma Interest Group. *Lung Cancer*, 14(1):1–12, 1996.

[30] R. M. Flores, M. Zakowski, E. Venkatraman, L. Krug, K. Rosenzweig, J. Dycoco, C. Lee, C. Yeoh, M. Bains, and V. Rusch. Prognostic factors in the treatment of malignant pleural mesothelioma at a large tertiary referral center. *J Thorac Oncol*, 2:957–965, 2007.

[31] J. E. Herndon, M. R. Green, A. P. Chahinian, J. M. Corson, Y. Suzuki, and N. J. Vogelzang. Factors predictive of survival among 337 patients with mesothelioma treated between 1984 and 1994 by the Cancer and Leukemia Group B. *Chest*, 113(3):723–731, 1998.

[32] D. Curran, T. Sahmoud, P. Therasse, J. van Meerbeeck, P. E. Postmus, and G. Giaccone. Prognostic factors in patients with pleural mesothelioma: The European Organization for Research and Treatment of Cancer experience. *J Clin Oncol*, 16:145–152, 1998.

[33] J. G. Edwards, K. R. Abrams, J. N. Leverment, T. J. Spyt, D. A. Waller, and K. J. O'Byrne.

Prognostic factors for malignant mesothelioma in 142 patients: Validation of CALGB and EORTC prognostic scoring systems. *Thorax*, 55:731–735, 2000.

[34] J. Francart, E. Vaes, S. Henrard, C. Legrand, P. Baas, R. Gaafar, J. P. van Meerbeeck, R. Sylvester, and A. Robert. A prognostic index for progression-free survival in malignant mesothelioma with application to the design of phase II trials: A combined analysis of 10 EORTC trials. *Eur J Cancer*, 45:2304–2311, 2009.

[35] J. Creaney, R. J. Francis, I. M. Dick, A. W. Musk, B. W. S. Robinson, M. J. Byrne, and A. Nowak. Serum soluble mesothelin concentrations in malignant pleural mesothelioma: Relationship to tumor volume, clinical stage and changes in tumor burden. *Clin Cancer Res*, 17(5):1181–1189, 2011.

[36] H. Pass, B. K. Temeck, K. Kranda, S. M. Steinberg, and I. R. Feuerstein. Preoperative tumor volume is associated with outcome in malignant pleural mesothelioma. *J Thorac Cardiovasc Surg*, 115(2):310–317, 1998.

[37] H. Y. Lee, S. H. Hyun, K. S. Lee, B.-T. Kim, J. Kim, Y. M. Shim, M.-J. Ahn, T. S. Kim, C. A. Yi, and M. J. Chung. Volume-based parameter of (18)F-FDG PET/CT in malignant pleural mesothelioma: Prediction of therapeutic response and prognostic implications. *Ann Surg Oncol*, 17(10):2787–2794, 2010.

[38] A. Nowak, R. J. Francis, M. J. Phillips, M. J. Millward, A. A. van der Schaaf, J. A. Boucek, A. W. Musk, M. J. McCoy, A. Segal, P. Robins, and M. J. Byrne. A novel prognostic model for malignant mesothelioma incorporating quantitative FDG-PET imaging with clinical parameters. *Clin Cancer Res*, 16(8):2409–2417, 2010.

[39] A. Nowak. CT, RECIST, and malignant pleural mesothelioma. *Lung Cancer*, 49 Suppl 1:S37–40, 2005.

[40] M. J. Ratain and S. G. Eckhardt. Phase II studies of modern drugs directed against new targets: If you are fazed, too, then resist RECIST. *J Clin Oncol*, 22(22):4442–4445, 2004.

[41] M. J. Ratain. Phase II oncology trials: Let's be positive. *Clin Cancer Res*, 11(16):5661–5662, 2005.

[42] A. B. Miller, B. Hoogstraten, M. Staquet, and A. Winkler. Reporting results of cancer treatment. *Cancer*, 47(1):207–214, 1981.

[43] P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst*, 92(3):205–216, 2000.

[44] L. C. Michaelis and M. J. Ratain. Measuring response in a post-RECIST world: From black

and white to shades of grey. *Nat Rev Cancer*, 6(5):409–414, 2006.

[45] C. C. Jaffe. Measures of response: RECIST, WHO, and new alternatives. *J Clin Oncol*, 24:3245–3251, 2006.

[46] P. Therasse, E. A. Eisenhauer, and J. Verweij. RECIST revisited: A review of validation studies on tumour assessment. *Eur J Cancer*, 42(8):1031–1039, 2006.

[47] M. J. Byrne and A. Nowak. Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. *Ann Oncol*, 15(2):257–260, 2004.

[48] R. J. van Klaveren, J. G. J. V. Aerts, H. de Bruin, G. Giaccone, C. Manegold, and J. P. van Meerbeeck. Inadequacy of the RECIST criteria for response evaluation in patients with malignant pleural mesothelioma. *Lung Cancer*, 43(1):63–69, 2004.

[49] S. G. Armato III and G. R. Oxnard. The radiologic measurement of mesothelioma. *Hematol Oncol Clin North Am*, 19(6):1053–1066, 2005.

[50] L. H. Schwartz, J. A. C. Colville, M. S. Ginsberg, L. Wang, M. Mazumdar, J. Kalaigian, H. Hricak, D. Ilson, and G. K. Schwartz. Measuring tumor response and shape change on CT: Esophageal cancer as a paradigm. *Ann Oncol*, 17(6):1018–1023, 2006.

[51] G. R. Oxnard, S. G. Armato III, and H. L. Kindler. Modeling of mesothelioma growth demonstrates weaknesses of current response criteria. *Lung Cancer*, 52(2):141–148, 2006.

[52] S. G. Armato III, G. R. Oxnard, H. MacMahon, N. J. Vogelzang, H. L. Kindler, M. Kocherginsky, and A. Starkey. Measurement of mesothelioma on thoracic CT scans: A comparison of manual and computer-assisted techniques. *Med Phys*, 31(5):1105–1115, 2004.

[53] S. G. Armato III, J. L. Ogarek, A. Starkey, N. J. Vogelzang, H. L. Kindler, M. Kocherginsky, and H. MacMahon. Variability in mesothelioma tumor response classification. *Am J Roentgenol*, 186(4):1000–1006, 2006.

[54] C. Plathow, M. Klopp, C. Thieke, F. Herth, A. Thomas, A. Schmaehl, I. Zuna, and H.-U. Kauczor. Therapy response in malignant pleural mesothelioma-role of MRI using RECIST, modified RECIST and volumetric approaches in comparison with CT. *Eur Radiol*, 18:1635–1643, 2008.

[55] R. J. Francis, M. J. Byrne, A. A. van der Schaaf, J. A. Boucek, A. Nowak, M. Phillips, R. Price, A. P. Patrikeos, A. W. Musk, and M. J. Millward. Early prediction of response to chemotherapy and survival in malignant pleural mesothelioma using a novel semiautomated 3-dimensional volume-based analysis of serial 18F-FDG PET scans. *J Nucl Med*, 48:1449–1458, 2007.

[56] P. Veit-Haibach, N. G. Schaefer, H. C. Steinert, J. D. Soyka, B. Seifert, and R. A. Stahel.

Combined FDG-PET/CT in response evaluation of malignant pleural mesothelioma. *Lung Cancer*, 67(3):311–317, 2010.

[57] T. Frauenfelder, M. Tutic, W. Weder, R. P. Götti, R. A. Stahel, B. Seifert, and I. Opitz. Volumetry: An alternative to assess therapy response for malignant pleural mesothelioma? *Eur Respir J*, 38(1):162–168, 2011.

[58] F. Liu, B. Zhao, L. M. Krug, N. M. Ishill, R. C. Lim, P. Guo, M. Gorski, R. Flores, C. S. Moskowitz, V. W. Rusch, and L. H. Schwartz. Assessment of therapy responses and prediction of survival in malignant pleural mesothelioma through computer-aided volumetric measurement on computed tomography scans. *J Thorac Oncol*, 5(6):879–884, 2010.

[59] J. A. Boucek, R. J. Francis, C. G. Jones, N. Khan, B. A. Turlach, and A. J. Green. Assessment of tumour response with (18)F-fluorodeoxyglucose positron emission tomography using three-dimensional measures compared to SUVmax–a phantom study. *Phys Med Biol*, 53(16):4213–4230, 2008.

[60] W. F. Sensakovic, S. G. Armato III, C. Straus, R. Y. Roberts, P. Caligiuri, A. Starkey, and H. L. Kindler. Computerized segmentation and measurement of malignant pleural mesothelioma. *Med Phys*, 38(1):238–244, 2011.

[61] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, NY, 2nd edition, 2010.

[62] D. G. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, New York, NY, 2nd ed edition, 2005.

[63] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 53(282):457–481, 1958.

[64] K. J. Jager, P. C. van Dijk, C. Zoccali, and F. W. Dekker. The analysis of survival data: The Kaplan-Meier method. *Kidney Int*, 74(5):560–565, 2008.

[65] J. M. Bland and D. G. Altman. The logrank test. *BMJ*, 328(7447):1073, 2004.

[66] D. Cox. Regression models and life tables (with discussion). *Jr Stat Soc B*, 34(2):187–220, 1972.

[67] M. Zhou. Understanding the Cox regression models with time-change covariates. *Am Stat*, 55:153–155, 2001.

[68] C. E. Metz. Basic principles of ROC analysis. *Semin Nucl Med*, 8:283–298, 1978.

[69] C. E. Metz. ROC methodology in radiologic imaging. *Invest Radiol*, 21:720–733, 1986.

[70] C. E. Metz. ROC analysis in medical imaging: A tutorial review of the literature. *Radiol*

Phys Technol, 1:2–12, 2008.

[71] F. E. Harrell Jr, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15(4):361–387, 1996.

[72] P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[73] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Stat Med*, 24(24):3927–3944, 2005.

[74] L. E. Chambless and G. Diao. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med*, 25(20):3474–3486, 2006.

[75] S.-H. Ma, H.-B. Le, B.-h. Jia, Z.-X. Wang, Z.-W. Xiao, X.-L. Cheng, W. Mei, M. Wu, Z.-G. Hu, and Y.-G. Li. Peripheral pulmonary nodules: Relationship between multi-slice spiral CT perfusion imaging and tumor angiogenesis and VEGF expression. *BMC Cancer*, 8:186, 2008.

[76] F. L. Giesel, H. Bischoff, H. von Tengg-Kobligk, M.-A. Weber, C. M. Zechmann, H.-U. Kauczor, and M. V. Knopp. Dynamic contrast-enhanced MRI of malignant pleural mesothelioma: A feasibility study of noninvasive assessment, therapeutic follow-up, and possible predictor of improved outcome. *Chest*, 129:1570–1576, 2006.

[77] K. A. Miles. Perfusion CT for the assessment of tumour vascularity: Which protocol? *Br J Radiol*, 76 Spec No 1:S36–42, 2003.

[78] A. R. Kambadakone and D. V. Sahani. Body perfusion CT: Technique, clinical applications, and advances. *Radiol Clin North Am*, 47:161–178, 2009.

[79] K. A. Miles and M. R. Griffiths. Perfusion CT: A worthwhile enhancement? *Br J Radiol*, 76:220–231, 2003.

[80] M. R. Meijerink, H. van Cruijsen, K. Hoekman, M. Kater, C. van Schaik, J. H. T. M. van Waesberghe, G. Giaccone, and R. A. Manoliu. The use of perfusion CT for the evaluation of therapy combining AZD2171 with gefitinib in cancer patients. *Eur Radiol*, 17:1700–1713, 2007.

[81] W. Bogdanich. The Radiation Boom; While Technology Surges, Radiation Safeguards Lag. *NY Times (Print)*, 2010.

[82] W. Bogdanich. The Radiation Boom; The Mark of an Overdose. *NY Times (Print)*, 2010.

[83] Y. Li, Z.-G. Yang, T.-W. Chen, Y.-P. Deng, J.-Q. Yu, and Z.-L. Li. Whole tumour perfusion of peripheral lung carcinoma: Evaluation with first-pass CT perfusion imaging at

64-detector row CT. *Clin Radiol*, 63:629–635, 2008.

[84] G. Brix, U. Lechel, G. Glatting, S. Ziegler, W. Munzing, S. Muller, and T. Beyer. Radiation exposure of patients undergoing whole-body dual-modality 18F-FDG PET/CT examinations. *J Nucl Med*, 46:608–613, 2005.

[85] W. D. Smith, R. C. Dutton, and N. T. Smith. A measure of association for assessing prediction accuracy that is a generalization of non-parametric ROC area. *Stat Med*, 15(11):1199–1215, 1996.

[86] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[87] F. E. Harrell Jr and with contributions from many other users. *Hmisc: Harrell Miscellaneous*. Programs available from http://CRAN.R-project.org/package=Hmisc, 2010.

[88] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Stat*, 37(1):36–48, 1983.

[89] D. D. Dorfman, K. S. Berbaum, and C. E. Metz. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol*, 27(9):723–731, 1992.

[90] C. A. Roe and C. E. Metz. Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Acad Radiol*, 4(8):587–600, 1997.

[91] K. R. Birchard, J. K. Hoang, J. E. Herndon, and E. F. Patz Jr. Early changes in tumor size in patients treated for advanced stage nonsmall cell lung cancer do not correlate with survival. *Cancer*, 115(3):581–586, 2009.

[92] W. F. Sensakovic, A. Starkey, R. Roberts, C. Straus, P. Caligiuri, M. Kocherginsky, and S. G. Armato III. The influence of initial outlines on manual segmentation. *Med Phys*, 37(5):2153–2158, 2010.

[93] G. Zalcman, J. Margery, A. Scherpereel, P. Astoul, I. Monnet, B. Milleron, C. Creveuli, M. P. Lebitasy, M. André, D. Moro-Sibilot, J. Mazieres, and on behalf of the French Intergroup of Thoracic Cancer Research (IFCT). IFCT-GFPC-0701 MAPS trial, a multicenter randomized phase II/III trial of pemetrexed-cisplatin with or without bevacizumab in patients with malignant pleural mesothelioma [abstract 7020]. *J Clin Oncol*, 28(15 Suppl), 2010.

[94] W. F. Sensakovic, A. Starkey, R. Y. Roberts, and S. G. Armato III. Discrete-space versus continuous-space lesion boundary and area definitions. *Med Phys*, 35(9):4070–4078, 2008.

[95] W. F. Sensakovic. *Computerized Segmentation and Measurement of Pleural Disease*. PhD

thesis, The University of Chicago, 2010.

[96] W. Hays. *Statistics*. Wadsworth Publishing, Belmont, CA, USA, 5th edition, 1994.

[97] J. L. Fleiss and P. E. Shrout. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, 43(2):259–262, 1978.

[98] J. L. Fleiss. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1999.

[99] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychol Bull*, 76(5):378–382, 1971.

[100] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[101] K. L. Gwet. *Handbook of Inter-Rater Reliability*. Advanced Analytics, LLC, Gaithersburg, MD, 2nd edition, 2010.

[102] J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327(8476):307–310, 1986.

[103] J. M. Bland and D. G. Altman. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat*, 17(4):571–582, 2007.

[104] S. R. Prasad, K. S. Jhaveri, S. Saini, P. F. Hahn, E. F. Halpern, and J. E. Sumner. CT tumor measurement for therapeutic response assessment: Comparison of unidimensional, bidimensional, and volumetric techniques–Initial observations. *Radiology*, 225(2):416–419, 2002.

[105] J. M. Boone. Radiological interpretation 2020: Toward quantitative image assessment. *Med Phys*, 34(11):4173–4179, 2007.

[106] E. Mehrara, E. Forssell-Aronsson, and P. Bernhardt. Objective assessment of tumour response to therapy based on tumour growth kinetics. *Br J Cancer*, 105(5):682–686, 2011.

[107] S. G. Armato III and W. F. Sensakovic. Automated lung segmentation for thoracic CT: Impact on computer-aided diagnosis. *Acad Radiol*, 11(9):1011–1021, 2004.

[108] W. F. Sensakovic, S. G. Armato III, A. Starkey, H. L. Kindler, and W. T. Vigneswaran. Quantitative measurement of lung reexpansion in malignant pleural mesothelioma patients undergoing pleurectomy/decortication. *Acad Radiol*, 18(3):294–298, 2011.

[109] A. P. Kiraly, S. Qing, and H. Shen. A novel visualization method for the ribs within chest volume data. In *Proceedings of SPIE, Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display*, page 614108, 2006.

[110] J. Staal, B. van Ginneken, and M. A. Viergever. Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data. *Med Image Anal*, 11(1):35–46, 2007.

[111] J. Lee and A. P. Reeves. Segmentation of individual ribs from low-dose chest CT. In *Proceedings of SPIE, Medical Imaging 2010: Computer-Aided Diagnosis*, page 76243J, 2010.

[112] S. Ramakrishnan, C. Alvino, L. Grady, and A. P. Kiraly. Automatic three-dimensional rib centerline extraction from CT scans for enhanced visualization and anatomical context. In *Proceedings of SPIE, Medical Imaging 2011: Image Processing*, page 7962104, 2011.

[113] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever. Multiscale vessel enhancement filtering. In W. M. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Interventation — MICCAI '98*, pages 130–137. Springer-Verlag, Berlin/Heidelberg, 1998.

[114] J. C. Russ. *The Image Processing Handbook*. CRC Press, Boca Raton, FL, 3rd edition, 1999.

[115] R. Beare and G. Lehmann. The watershed transform in ITK–Discussion and new developments. *Insight Journal*, 2006.

[116] G. T. Herman, J. Zheng, and C. A. Bucholtz. Shape-based interpolation. *IEEE Comput Grap Appl*, 12(3):69–79, 1992.

[117] H. Akaike. A new look at the statistical model identification. *IEEE Trans Automat Contr*, 19(6):716–723, 1974.

[118] M. Yi, E. A. Mittendorf, J. N. Cormier, T. A. Buchholz, K. Bilimoria, A. A. Sahin, G. N. Hortobagyi, A. M. Gonzalez-Angulo, S. Luo, A. U. Buzdar, J. R. Crow, H. M. Kuerer, and K. K. Hunt. Novel staging system for predicting disease-specific survival in patients with breast cancer treated with surgery as the first intervention: Time to modify the current American Joint Committee on Cancer staging system. *J Clin Oncol*, 29(35):4654–4661, 2011.

[119] T. Hothorn and A. Zeileis. Generalized maximally selected statistics. *Biometrics*, 64(4):1263–1269, 2008.

[120] B. Lausen and M. Schumacher. Maximally selected rank statistics. *Biometrics*, 48(1):73–85, 1992.

[121] S. G. Armato III, G. R. Oxnard, M. Kocherginsky, N. J. Vogelzang, H. L. Kindler, and H. MacMahon. Evaluation of semiautomated measurements of mesothelioma tumor thickness on CT scans. *Acad Radiol*, 12(10):1301–1309, 2005.

[122] H. L. Kindler. Systemic treatments for mesothelioma: Standard and novel. *Curr Treat Options Oncol*, 9(2-3):171–179, 2008.

[123] S. K. Sha, T. Sato, H. Kobayashi, M. Ishigaki, S. Yamamoto, H. Sato, A. Takada, S. Nakajyo, Y. Mochizuki, J. M. Friedman, F. C. Cheng, T. Okura, R. Kimura, D. W. Kufe, D. D. VonHoff, and T. Kawabe. Cell cycle phenotype-based optimization of G2-abrogating peptides yields CBP501 with a unique mechanism of action at the G2 checkpoint. *Mol Cancer Ther*, 6(1):147–153, 2007.

[124] M. Wintermark, W. S. Smith, N. U. Ko, M. Quist, P. Schnyder, and W. P. Dillon. Dynamic perfusion CT: Optimizing the temporal resolution and contrast volume for calculation of perfusion CT parameters in stroke patients. *Am J Neuroradiol*, 25:720–729, 2004.

[125] J. A. Christner, J. M. Kofler, and C. H. McCollough. Estimating effective dose for CT using dose-length product compared with using organ doses: Consequences of adopting International Commission on Radiological Protection publication 103 or dual-energy scanning. *Am J Roentgenol*, 194(4):881–889, 2010.

[126] C. H. McCollough, S. Leng, L. Yu, D. D. Cody, J. M. Boone, and M. F. McNitt-Gray. CT dose index and patient dose: They are not the same thing. *Radiology*, 259(2):311–316, 2011.

[127] W. Huda, K. M. Ogden, and M. R. Khorasani. Converting dose-length product to effective dose at CT. *Radiology*, 248(3):995–1003, 2008.

[128] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Phys Med Biol*, 46(3):R1–45, 2001.

[129] J. Piper, Y. Ikeda, Y. Fujisawa, Y. Ohno, T. Yoshikawa, A. O'Neil, and I. Poole. Objective evaluation of the correction by non-rigid registration of abdominal organ motion in low-dose 4D dynamic contrast-enhanced CT. *Phys Med Biol*, 57(6):1701–1715, 2012.

[130] A. A. Isola, H. Schmitt, U. van Stevendaal, P. G. Begemann, P. Coulon, L. Boussel, and M. Grass. Image registration and analysis for quantitative myocardial perfusion: Application to dynamic circular cardiac CT. *Phys Med Biol*, 56(18):5925–5947, 2011.

[131] A. Chandler, W. Wei, D. H. Herron, E. F. Anderson, V. E. Johnson, and C. S. Ng. Semiautomated motion correction of tumors in lung CT-perfusion studies. *Acad Radiol*, 18(3):286–293, 2011.

[132] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung. Validation of an accelerated "demons" algorithm for deformable image registration in radiation therapy. *Phys Med Biol*, 50(12):2887–2905, 2005.

[133] J. Thirion. Image matching as a diffusion process: An analogy with Maxwell's demons.

Med Image Anal, 2(3):243–260, 1998.

[134] G. C. Sharp, N. Kandasamy, H. Singh, and M. Folkert. GPU-based streaming architectures for fast cone-beam CT image reconstruction and demons deformable registration. *Phys Med Biol*, 52(19):5771–5783, 2007.

[135] K. Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. de Bruijne, X. Han, M. P. Heinrich, J. A. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. R. McClelland, S. Ourselin, S. E. A. Muenzing, M. A. Viergever, D. De Nigris, D. L. Collins, T. Arbel, M. Peroni, R. Li, G. C. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. C. Gee, M. Staring, S. Klein, B. C. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. P. W. Pluim. Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge. *IEEE Trans Med Imaging*, 30(11):1901–1920, 2011.

[136] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[137] N. A. Mullani and K. L. Gould. First-pass measurements of regional blood flow with external detectors. *J Nucl Med*, 24(7):577–581, 1983.

[138] M. J. Blomley, R. Coulden, C. Bufkin, M. J. Lipton, and P. Dawson. Contrast bolus dynamic computed tomography for the measurement of solid organ perfusion. *Invest Radiol*, 28(Suppl 5):S72–S78, 1993.

[139] M. J. Blomley, R. Coulden, P. Dawson, M. Kormano, P. Donlan, C. Bufkin, and M. J. Lipton. Liver perfusion studied with ultrafast CT. *J Comput Assist Tomogr*, 19(3):424–433, 1995.

[140] M. M. Koenig, E. E. Klotz, B. B. Luka, D. J. D. Venderink, J. F. J. Spittler, and L. L. Heuser. Perfusion CT of the brain: Diagnostic approach for early detection of ischemic stroke. *Radiology*, 209(1):85–93, 1998.

[141] L. Axel. Cerebral blood flow determination by rapid-sequence computed tomography: Theoretical analysis. *Radiology*, 137(3):679–686, 1980.

[142] D. J. J. Wang, X. Bi, B. B. Avants, T. Meng, S. Zuehlsdorff, and J. A. Detre. Estimation of perfusion and arterial transit time in myocardium using free-breathing myocardial arterial spin labeling with navigator-echo. *Magn Reson Med*, 64(5):1289–1295, 2010.

[143] P. S. Tofts, G. Brix, D. L. Buckley, J. L. Evelhoch, E. Henderson, M. V. Knopp, H. B. Larsson, T. Y. Lee, N. A. Mayr, G. J. Parker, R. E. Port, J. Taylor, and R. M. Weisskoff. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a

diffusable tracer: Standardized quantities and symbols. *J Magn Reson Imaging*, 10(3):223–232, 1999.

[144] T. G. Karrison, M. L. Maitland, W. M. Stadler, and M. J. Ratain. Design of phase II cancer trials using a continuous endpoint of change in tumor size: Application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J Natl Cancer Inst*, 99:1455–1461, 2007.