Volumetric Assessment of Hepatocellular Carcinoma Response to Treatment Using a Random Forest-Based Automated Segmentation Protocol

Kareem Ahmed, MD[1]

David Fuentes, PhD[1]

J.S. Lin, PhD[1]

Reham Ali, MD[2]

Ahmed O. Kaseb, MD[2]

Manal Hassan, MD, PhD

Janio Szklaruk, MD, PhD

John D. Hazle, MD[1]

Aliya Qayyum, MD[4]

Khaled M. Elsayes, MD[4]

Departments of [1]Imaging Physics, [2]Gastrointestinal Oncology, [3]Biostatistics, and [4]Diagnostic Radiology,

The University of Texas MD Anderson Cancer Center,

Houston, TX 77030, USA

Corresponding author: Khaled M. Elsayes, MD, Professor, Department of Radiology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030; phone: 713-745-3025; fax: 713-794-4379; email: kmelsayes@mdanderson.org.

K. Ahmed,[1] D. Fuentes,[1] J.S. Lin,[1] R. Ali,[2] A.O. Kaseb,[2] W. Wei,[3] J.D. Hazle,[1] A. Qayyum,[4]

K.M. Elsayes[4]

Abstract

***Purpose*:** To determine whether machine learning based 3D volumetric quantification of contrast enhancing and non-enhancing portions of hepatocellular carcinomas (HCC) can be used as an imaging biomarker to predict the response to treatment by Sorafenib or TACE.

Methods: A training database of manually defined background liver, enhancing tumor, and non-enhancing tumor tissue was established using pre- and the first post-therapy multiphasic CT images from 30 patients. For each patient, intensity- and geometry-based feature images were generated from registered pre-contrast and tri-phasic CT datasets and used as the input for the Random Forest (RF)-based classifier. Leave-one-out cross validation was applied to every permutation of the training data subset to quantify the classifier accuracy. Volume changes of enhancing and non-enhancing tumor portions (Volumetric RECIST [vRECIST]) was calculated and used to classify patients as responders and non-responders using the volume formula of a sphere. Kaplan-Meier survival analysis with log-rank test was performed to quantitatively compare the stratification of responders and non-responders calculated with both manual and automated methods. A Cox proportional hazard ratio model was used to confirm survival results.

Results: The manual vRECIST of enhancing (hazard ratio, 0.16; 95% CI: 0.28, 0.93; P = 0.042) and non-enhancing (hazard ratio, 0.057; 95% CI: 0.004, 0.678; P = 0.023) portions of HCC tumor lesions showed significant difference in TTP of target lesions between responders and non-responders. We observed an overall classification accuracy of median Dice similarity coefficient (DSC) of 0.81 prior to and 0.66 after therapy for the enhancing tumor labels and 0.66 prior to and 0.82 after therapy for the non-enhancing tumor label.

Conclusion: In this study, we determined that automated 3D volumetric analysis was a reliable and reproducible method of monitoring HCC response to treatment with Sorafenib or TACE. With respect to TTP and clinical outcome, the information extracted using automated methods of quantitative liver lesion volumetrics was statistically equivalent to labor-intensive manually define volumes.

Keywords

Image Segmentation · HCC · Random Forest · TACE · Sorafenib

1 Introduction

More than 748,000 new cases of hepatocellular carcinoma (HCC) are diagnosed globally per year. HCC is ranked third among the leading causes of cancer-related death worldwide [1]. Multiple treatment options are available for unresectable HCC; Sorafenib is an oral multikinase inhibitor that interferes with tumor cell signaling pathways resulting in reduction of tumor neoangiogenesis and stimulation of apoptosis.[2, 3] Macroscopically, it reduces tumor vascularity and induces intra-lesion necrosis. [4] Transcatheter arterial chemoembolization (TACE) is a local-regional treatment that selectively delivers high-concentration Doxorubicin to targeted tumor lesions.

Tumor response assessment is crucial for the evaluation of HCC therapy. Current guidelines established by The European Association for the Study of the Liver (EASL) and The American Association for the Study of Liver Diseases (AASLD) criteria, also known as modified-RECIST (mRECIST) criteria, adopt the uni- and bidimensional measurements of enhancing portions of target tumor lesions in single representative axial cut to assess therapeutic effect. (Enhancing tumor is defined as “uptake of contrast agent in the arterial phase of dynamic contrast CT or MRI”, while non-enhancing tumor is defined as “regions of no enhancement within HCC on arterial phase images”) [5-7]. Multiple studies have highlighted the limitation of this method compared to volumetric analysis when applied to HCC because the treatment-induced changes are inhomogeneous; thus, not only is single slice selection less representative of the actual tumor burden, it is subject to intra-observer and inter-observer variability;

meanwhile, volumetric analysis can overcome such limitations and provide reliable and reproducible alternative, with excellent intra-observer and inter-observer agreement[8-11]. Manual and semi-automated volumetric analysis is time consuming and remains observer dependent. Therefore, unsupervised (user independent) HCC segmentation can be a reproducible and time efficient alternative. Since the HU values of the liver within CT images are typically very similar to those of the surrounding tissue, unsupervised identification of the unpredictable appearance of the tumor is one of the most challenging tasks in medical image analysis[12]. For this reason, tissue segmentation based on HU intensity values alone is not feasible, and numerous approaches have incorporated various levels of prior information based on geometrical and anatomical considerations [13-15]. In this study, we evaluate the role of 3D volumetric quantification of enhancing and non-enhancing portions of HCC as an imaging biomarker to predict the response to treatment by Sorafenib or TACE. Additionally, we compare an automated RF-based volumetric quantification method to the labor-intensive, operator-dependent manual approach.

Crowdsourcing-based algorithmic challenges in medical image segmentation [12, 16], combined with software reproducibility efforts [17, 18], motivated the technical aspects of the present study. Machine learning techniques based on random forest classification algorithms are adapted to quantify viable and non-viable tumor volumes in HCC patient populations for this application. Quantitative image features used in the analysis are derived from both geometrical and intensity models of the imaging.

2 Materials and Methods

2.1 Study Cohort

This IRB approved retrospective, single-institution study included 30 patients diagnosed with HCC, including 23 men (mean age, 70.5; range, 55-86) and 7 women (mean age, 74.5; range, 56-93). Patients received Sorafenib (13 patients) or TACE (17 patients) as the first line of treatment at our institution between 8/2008 and 4/2014. The inclusion criteria were:

Unresectable HCC, Child-Pugh score A or B, Sorafenib or TACE as the first line of treatment, and multiphasic contrast-enhanced CT at baseline and a minimum of 4 weeks after treatment.

The baseline CT was performed at mean time interval of 1 week before the first dose of Sorafenib and 2 weeks before the first session of TACE. The first follow-up CT was at mean time of 9 weeks and 10 weeks after the start of Sorafenib or the first TACE procedure, respectively. Patients treated with Sorafenib received a standard dose of 400 mg twice daily. Treatment interruption or dose reduction was allowed in cases of adverse drug reactions. Patients treated with TACE received Doxorubicin in LC beads. The study endpoint was time to progression (TTP) of the target lesion. Table 1 shows patients' demographic data and clinical profiles.

2.2 CT Imaging Technique

All patients underwent multiphasic contrast-enhanced CT of the abdomen (4-,16-, or 64-MDCT Light-Speed, GE Healthcare). The liver protocol was used in all studies (scanning was performed during the arterial phase (bolus tracking) 17 seconds after peak enhancement of the aorta after contrast injection, the porto-venous phase at 60 seconds, and the delayed phase approximately 3-5 minutes). The injection rate was 3-5 mL/s, image reconstruction thickness was 2.5 and 5-mm. A total of 240 CT image series (60 CT studies, each with pre-contrast, arterial, porto-venous, and delayed phases) from 30 patients were examined. An example image set from a single study is shown in Figure 1.

2.3 Data Curation and Training Data

Manual segmentations were performed using semi-automatic segmentation tools available in AMIRA [1] to delineate the (1) liver, (2) enhancing tumor, and (3) non-enhancing tumor on the porto-venous phase images from the 30 patients. **A radiology resident and a professor with 20 years of experience** in abdominal imaging performed the manual segmentation. This resulted in 3 tissue labels: label 1 for the background liver parenchyma, 2 for the enhancing tumor region, and 3 for the non-enhancing tumor

[1]FEI TM, http://www.fei.com/software/amira-3d-for-life-sciences/

region, before and after treatment. For a given patient study, pre-contrast, arterial, and delayed contrast phase images were registered to the porto-venous contrast phase images. This allowed the manual labels to be applied to all images in the set. The 30 training sets provided the gold standard reference for the classifier prediction. All DICOM images were converted to NifTi format to preserve the orientation information for data processing.

2.4 Image Features

Prior to image feature extraction, the registered and masked tri-phasic and pre-contrast CT datasets, shown in Figure 1, were processed with a total variation denoising filter to denoise the image while preserving the image boundaries [19]. A comprehensive list of the image features considered is shown in Table 2. A set of 105 total images features was extracted and consisted of the statistical summary of the HU intensity values within a pixel neighborhood (mean, standard deviation, and Skewness) for pixel radius of 1, 3, and 5. Mixture model probabilities and geometrical descriptors of each label were included as image features [20]. The geometrical descriptors include the distance to the tumor core, elongation, eccentricity, volume, and surface-to-volume ratio of the mixture model segmentation label. The distance to the liver mask was also considered. Distance features were computed using the Maurer distance transformation [21].

2.5 Random Forest (RF) Decision Trees

In this application, we focused on the use of random forest-based methods for the classification of diseased and background liver tissue. Subsequent volumetrics provide response criteria to evaluate response to HCC treatment [17]. The input for our random forest classification method was the set of image features shown in Table 2, that were derived from (1) pre-contrast and tri-phasic contrast enhanced-CT of the abdomen in the (2) arterial phase, (3) washout in the porto-venous, and (4) the delayed phases. During ‘training’, the classifier is constructed as a collection of decision trees that are calibrated to a random subset of the manually labeled training data [22]. A key principle of random

forest methods is that random features selection testing as well as training on random subsets of data, decrease the correlation between outputs of different decision trees while simultaneously improving the overall performance of the decision forest [22, 23]. Specifically, in this application, the number of trees was set to 500; for each tree, 2000 voxel samples were used from each label, representing the enhancing tumor, non-enhancing tumor, and background liver within the training data.

Given a calibrated or trained model, as shown in Figure 2, model prediction involves pixel-wise processing of image features by each decision tree. These features are processed by performing a series of binary tests along each internal node from the root to a leaf. Decision thresholds at each binary test are identified as the quantitative feature image value that best separates the collection of training dataset values with respect to the Gini impurity[24]. Each decision tree is used to determine the tissue classification, based on the classification (normal liver, enhancing tumor, or non-enhancing tumor) assigned to the leaf during the training stage. The classification probability is estimated as the fraction of classification from all trees. The final classification is used as the maximum probability.

2.6 Cross Validation

Leave-one-out cross validation was performed to quantify the prediction accuracy of the model. As the name implies, training algorithms are applied to available subsets of the patient cohort. This process closely simulates the clinical scenario in which all datasets are included to calibrate the model. Datasets that are not included in the calibration provide an independent test case for the quantitative evaluation of prediction accuracy. Dice similarity coefficient (DSC) was used as the quantifying overlap of the manual label and was used to predict classification. The overlap percentage defined as the intersection between the 2 contours divided by the union of the 2 contours, is calculated for each lesion. A model is considered accurate if the overlap percentage was .70 or higher [25].

2.7 Statistical Analysis

Therapy response was estimated as percent change in volume of each of enhancing and non-enhancing tumor portions between baseline and follow up CT. Results from manual and RF-based automated segmentation were compared. Time-to-progression (TTP) was calculated from the date of first dose of Sorafenib or first session of TACE till the date of progression of the target lesions according to the mRECIST and EASL guidelines currently used in our institution.

Kaplan-Meier survival analysis was conducted between responders and non-responders according to enhancing and according to non-enhancing tumor changes. Patients were censored at the time they switched to different treatment modality, were lost to follow up, or had liver transplantation . Results were analyzed with the log-rank test. Multivariate Cox regression was used to assess the effect of volume change on suvival.

Patients were stratified for Kaplan-Meier analysis according to change in volume of enhancing and non-enhancing tumor portions and according to the method of volumetric assessment, manual or automated. Four groups were established, change in enhancing tumor measured by manual segmentation, change in enhancing tumor calculated by the RF-automated method, change in non-enhancing tumor measured manually, and change in non-enhancing tumor calculated by the RF-automated method. In each group, patients were classified as responders and nonresponders based on the percent of volume change. A change in volume for enhancing (65% decrease) [11]and non-enhancing (35% increase) tumor was considered as responders. To our knowledge, there are no available guidelines for non-enhancing tumor portions. We propose the complementary of 65% which is 35% increase in volume of necrosis as cut off value for response. Thus, patients that showed more than 35% increase in volume of non-enhancing portion of the tumor were considered responders. Statistical analysis was performed using SPSS software (IBM SPSS Statistics, version 23).

3 Results

3.1 Prediction Accuracy

Representative image segmentations and feature images are shown in Figure 3. Manual segmentations are shown as "Truth" and serve as both (i) the gold standard reference for training the random forest model and (ii) an independent reference for evaluating prediction accuracy. The "RF Model" represents tissue classification using the random forest model, which was trained independently from the reference dataset within the cross validation analysis.

The mean, median, standard deviation summary statistics of the computed DSCs are provided in Table 3. The overall median prediction accuracy in classifying non-enhancing tissue was DSC=.662 and DSC=.817 before and after treatment, respectively. Similarly, the accuracy in classifying enhancing tumor was DSC=.809 and DSC=.664 before and after treatment; respectively. Figure 4 provides a comprehensive overview of the prediction accuracy for each patient and treatment modality in the cohort. Prediction accuracy is presented in terms of the DSC on the right axis. For reference, the corresponding volumes are plotted on the left axis.

3.2 Volumetric Analysis

Volumetric analysis by manual segmentation showed that the volume of enhancing tumor regions ranged from 3080.25 to 1,802,340 mm^3 (mean, 262,930.5) and from 0 to 3,487,530 mm^3 (mean, 309,293.3) at baseline and follow up, respectively. The volume of intratumoral necrosis ranged from 0 to 143,503 mm^3 (mean, 20,369.8) and from 0 to 874,461 mm^3 (mean, 67,926.3) at baseline and follow up, respectively. 9 patients showed more than 65% decrease in volume of enhancing tumor on follow up CT and are considered responders according to enhancing tumor change with treatment. On the other hand, 15 patients showed more than 35% increase in volume of non-enhancing tumor on follow up CT and are considered responders according to non-enhancing tumor change with treatment.

3.3 Survival Analysis

During the observational period, the target lesions of 22 (73.3%) patients progressed, 8 patients (26.7%) were censored because of intolerance to Sorafenib (n=1), switched to radiofrequency ablation (n=1) or to Sorafenib (n=3) after one or more TACE sessions, lost to follow up (n=2) or underwent liver transplantation (n=1). TTP of responders and nonresponders according to volume change of enhancing and non-enhancing tumor by manual and automated segmentation was compared using Kaplan-Meier survival curves shown in (Figures 5,6). Log-rank test revealed that the volume change of both enhancing and non-enhancing tumor can significantly classify responding and nonresponding lesions with accurate correlation with their time to progression; P value = 0.001 and 0.0001 for manually segemented enhancing and non-enhancing tumor, respectively, and is 0.02 and 0.66 for automatically segmented enhancing and non-enhancing tumor, respectively. Multivariate cox-regression analysis confirmed volume change of enhancing and non-enhancing portion of HCC lesion as predictors of survival (Table 4).

4 Discussion and Conclusion

The main outcome of this study is that 3D volumetric quantification of enhancing and non-enhancing tumor portions of HCC can be used as an imaging biomarker to predict the response to treatment. Additionally, our proposed RF-based automated method is statistically equivalent to labor-intensive manually labeled data with the advantage of elimination of inter- and intraobserver variability. Investigators have highlighted the superiority of volumetric analysis over the 1D and 2D measurement in single slice approach adopted by EASL and mRECIST guidelines. Galizia et al showed that 3D volumetric analysis was more reproducible than the corresponding 2D approach in a study in 29 patients with HCC treated with Y90 radioembilization therapy[8]. Good inter- and intra-observer reproducibility of semi-automated segmentation of liver tumor lesions has also been demonstrated by Monsky et al in 29 patients with HCC or liver metastasis[9]. Feasibility of volumetric approach was shown by Lin et al in a

study on 17 patient with HCC undergoing TACE[10]. Tacher et al pointed out that 3D tumor assessment methods (vRECIST and qEASL) are more accurate than EASL and mRECIST when predict HCC patient survival after the first TACE[11]. These results agree with the expected response of HCC to treatment modalities which target the tumor blood supply, such as Sorafenib and TACE. This results in shrinkage of the enhancing tumor and expansion of non-enhancing regions. This mechanism of action is reflected on the clinical outcomes in the form of longer TTP of the target lesion.

A Kruskal-Wallis test revealed that mixture model intensity classifications were the most significant for automated tumor segmentation by the RF model. Such findings agree with the radiologists' intuition, which depends almost entirely on discrepancies in intensity between the tumor and background liver throughout different contrast phases of the multiphasic CT protocol. This explains why the RF model better classified enhancing tumor in pre-therapy images while necrosis prediction was more accurate in post-therapy images. Since Sorafenib and TACE interfere with tumor blood supply, tumor is expected to be more enhancing in baseline images compared to post-therapy images, which is translated into more discrepancy in HU units between enhancing tumor, non-enhancing tumor and background liver and in turn, better delineation of the enhancing tumor. On the other hand, post-therapy images show areas of decreased enhancement and larger areas of no enhancement, consequently more discrepancy in HU between non-enhancing areas and background liver and in turn better delineation.

Enhancing Errors between manual and automated segmentation approaches were noticed mostly at the periphery of the tumor lesion; particularly in small lesions. This can be because intensity values at the periphery were close to that of the surrounding liver. In addition, registration accuracy may account for errors at the periphery of the tumor lesion as well. Decoupling and quantifying the effect of registration techniques on the resulting segmentation accuracy is the topic of ongoing studies.

The limitations of this study included small sample size due to our strict inclusion criteria. One other limitation is that the liver had to be manually masked during the preprocessing pipeline; this step has rendered the entire process semi-automated rather than fully automated. We aim to implement software

with automated liver segmentation capabilities in future studies to achieve a fully automated protocol for HCC volumetric analysis.

In conclusion, automated 3D volumetric analysis was a reliable and reproducible method of monitoring HCC response to treatment with Sorafenib or TACE. With respect to TTP and clinical outcome, the information extracted using automated methods of quantitative liver lesion volumetrics was statistically equivalent to labor-intensive manually labeled data.

5 Acknowledgments

This work was supported in part by the O'Donnell Foundation and NIH DP2OD007044-01S1 funding mechanisms. The authors also thank the open source communities ITK [26], ANTs[27], and itk-SNAP [25] for providing enabling software for image processing and visualization. The authors declare that they have no conflict of interest.

References

1. Yang, J.D. and L.R. Roberts, *Epidemiology and management of hepatocellular carcinoma*. Infectious disease clinics of North America, 2010. **24**(4): p. 899-919.
2. Wilhelm, S.M., et al., *BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis*. Cancer research, 2004. **64**(19): p. 7099-7109.
3. Chang, Y.S., et al., *Sorafenib (BAY 43-9006) inhibits tumor growth and vascularization and induces tumor apoptosis and hypoxia in RCC xenograft models*. Cancer Chemother Pharmacol, 2007. **59**(5): p. 561-74.
4. Liccioni, A., M. Reig, and J. Bruix, *Treatment of hepatocellular carcinoma*. Dig Dis, 2014. **32**(5): p. 554-63.
5. Bruix, J., M. Sherman, and D. American Association for the Study of Liver, *Management of hepatocellular carcinoma: an update*. Hepatology, 2011. **53**(3): p. 1020-2.
6. Lencioni, R. and J.M. Llovet. *Modified RECIST (mRECIST) assessment for hepatocellular carcinoma*. in *Seminars in liver disease*. 2010.
7. Bruix, J., et al., *Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver*. J Hepatol, 2001. **35**(3): p. 421-30.
8. Galizia, M.S., et al., *MDCT necrosis quantification in the assessment of hepatocellular carcinoma response to yttrium 90 radioembolization therapy: comparison of two-dimensional and volumetric techniques*. Acad Radiol, 2012. **19**(1): p. 48-54.
9. Monsky, W.L., et al., *Semiautomated segmentation for volumetric analysis of intratumoral ethiodol uptake and subsequent tumor necrosis after chemoembolization*. AJR Am J Roentgenol, 2010. **195**(5): p. 1220-30.
10. Lin, M., et al., *Quantitative and VolumetricEuropean Association for the Study of the Liver and Response Evaluation Criteria in Solid Tumors Measurements: Feasibility of a Semiautomated Software Method to Assess Tumor Response after Transcatheter Arterial Chemoembolization*. Journal of Vascular and Interventional Radiology, 2012. **23**(12): p. 1629-1637.
11. Tacher, V., et al., *Comparison of Existing Response Criteria in Patients with Hepatocellular Carcinoma Treated with Transarterial Chemoembolization Using a 3D Quantitative Approach*. Radiology, 2016. **278**(1): p. 275-84.
12. Heimann, T., et al., *Comparison and evaluation of methods for liver segmentation from CT datasets*. Medical Imaging, IEEE Transactions on, 2009. **28**(8): p. 1251-1265.
13. Lamecker, H., T. Lange, and M. Seebass, *Segmentation of the liver using a 3D statistical shape model*. 2004:

Citeseer.
14. Park, H., P.H. Bland, and C.R. Meyer, *Construction of an abdominal probabilistic atlas and its application in segmentation*. Medical Imaging, IEEE Transactions on, 2003. **22**(4): p. 483-492.
15. Soler, L., et al., *Fully automatic anatomical, pathological, and functional segmentation from CT scans for hepatic surgery*. Computer Aided Surgery, 2001. **6**(3): p. 131-142.
16. Menze, B., M. Reyes, and K. Van Leemput, *The Multimodal Brain TumorImage Segmentation Benchmark (BRATS)*. 2014.
17. Tustison, N.J., et al., *Optimal Symmetric Multimodal Templates and Concatenated Random Forests for Supervised Brain Tumor Segmentation (Simplified) with ANTsR*. Neuroinformatics, 2014: p. 1-17.
18. McCormick, M., et al., *ITK: enabling reproducible research and open science*. Frontiers in neuroinformatics, 2014. **8**.
19. Wolf, I., et al., *The medical imaging interaction toolkit*. Medical image analysis, 2005. **9**(6): p. 594-604.
20. Schmitt, P., et al., *Effects of slice thickness and head rotation when measuring glioma sizes on MRI: in support of volume segmentation versus two largest diameters methods*. Journal of neuro-oncology, 2013. **112**(2): p. 165-172.
21. Maurer Jr, C.R., R. Qi, and V. Raghavan, *A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003. **25**(2): p. 265-270.
22. Breiman, L., *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*. Statistical Science, 2001. **16**(3): p. 199-231.
23. Mahapatra, D., *Analyzing training information from random forests for improved image segmentation*. Image Processing, IEEE Transactions on, 2014. **23**(4): p. 1504-1512.
24. Liaw, A. and M. Wiener, *Classification and regression by randomForest*. R news, 2002. **2**(3): p. 18-22.
25. Zou, K.H., et al., *Three validation metrics for automated probabilistic image segmentation of brain tumours*. Statistics in medicine, 2004. **23**(8): p. 1259-1282.
26. Ibanez, L., et al., *The ITK software guide*. 2003.
27. Avants, B.B., et al., *A reproducible evaluation of ANTs similarity metric performance in brain image registration*. Neuroimage, 2011. **54**(3): p. 2033-44.

Table 1: Patient Population.

Baseline Patient Characteristics	
Patient Characteristic (n=30)	**Finding**
Age	
Mean	68
<60 y	7
>60 y	23
Sex	
Male	23
Female	7
Cirrhosis	
Present	23
Absent	7
Child-Pugh stage	
A	29
B	1

C	0
BCLC stage	
A	4
B	4
C or D	22
Alpha-Fetoprotein level	
Mean (ng/mL)	2226.963
<=400 ng/mL	23
>400 ng/mL	7
Tumor Nodularity	
Uninodular	11
Multinodular	19
Dose	
Sorafenib (mg) (Mean)	37953.8
Doxorubicin (mg) (Mean)	56.14

Table 2: Feature images used to construct the random forest model.

Intensity modeling and connected component geometry		
Feature	Number	Motivation
Pr (Background liver)	1 per modality	Background liver intensity
Pr (Enhancing tumor)	1 per modality	Enhancing tumor intensity
Pr (Necrotic tumor)	1 per modality	Necrotic tumor intensity
Elongation	1 per modality	Anisotropic components
Eccentricity	1 per modality	Anisotropic components
Volume	1 per modality	Small, isolated components
Distance to tumor core	1 per modality	Proximity to tumor core
Volume/surface area	1 per modality	Anisotropic components
Neighborhood first-order statistics		
Feature	Number	Motivation
Mean (radius=1,3.5)	1 per modality	Liver & tumor intensity
Std. dcv. (radius=1.3,5)	1 per modality	Liver & tumor intensity
Skewness (radius=1,3,5)	1 per modality	Liver & tumor intensity
Intensity gradient	1	Liver & tumor intensity
Pre,Art difference	1	Isolated tumor tissue
Art,Pori-venous difference	1	Isolated tumor tissue
Port-venous,Del difference	1	Isolated tumor tissue
Liver mask coordinate system		
Liver mask coordinate system		
Feature	Number	Motivation
Subject distance	1	Peripheral Vs Central

Table 3: Stat Summary of Error measurements.

Median/Mean (Std)	Overall enhancing	Overall non-enhancing
Prior	0.809/0.709 (0.035)	0.662/0.652 (0.041)
After	0.664/0.644 (0.045)	0.817/0.783 (0.040)

Table 4: Cox hazard regression model for the effect of change in volume enhancing and non-enhancing tumor on survival

Response	Hazard Ratio	P-value	95% Confidence Interval	
Change in volume of enhancing tumor	**0.163**	**0.042**	**0.028**	**0.937**
Change in volume of non-enhancing tumor	**0.057**	**0.023**	**0.005**	**0.678**

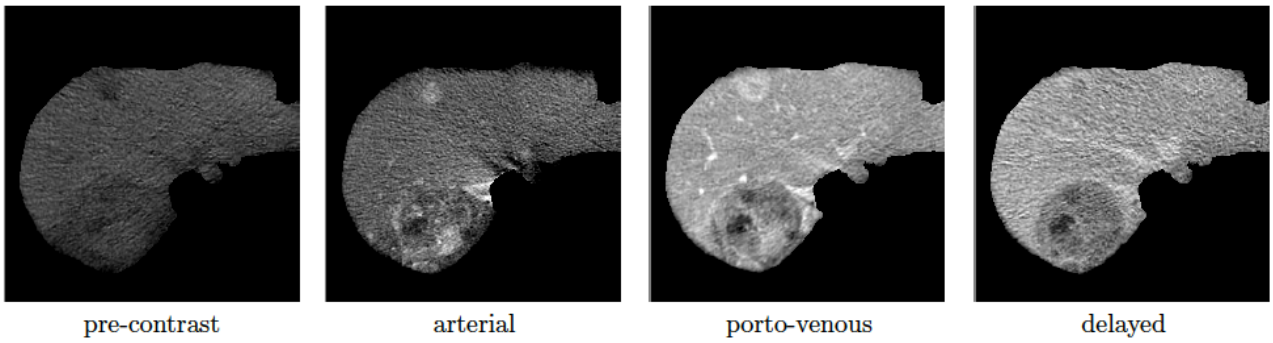


Figure 1: Original Images. Each study consists of 0.images from the pre-contrast, arterial, porto-venous, delayed phase shown.

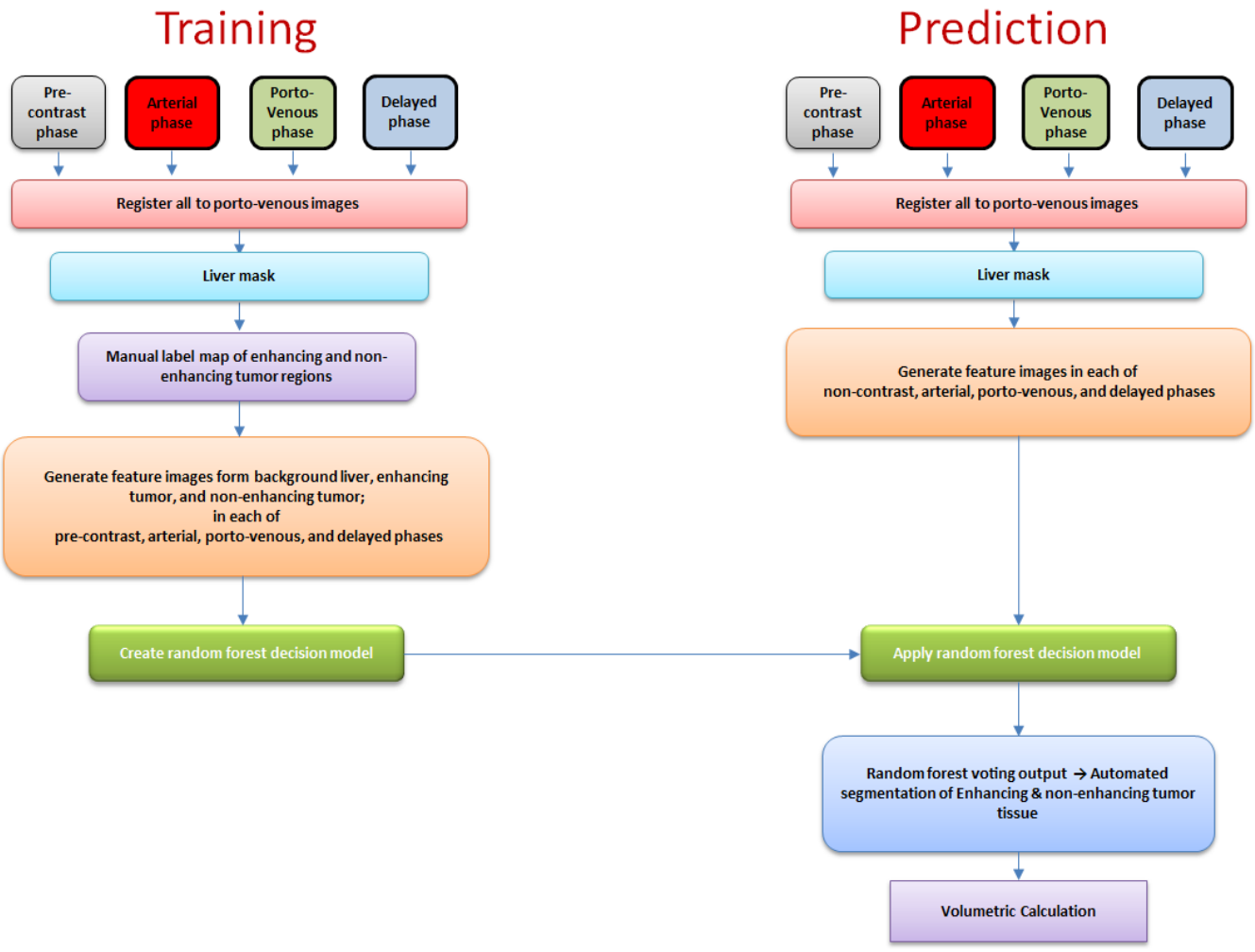


Figure 2: Workflow of the automated segmentation protocol. Each imaging set is registered to the respective porto-venous phase image and masked for processing. During training, quantitative image

features thresholds are identified that discriminate the manually labeled (i) normal tissue, (ii) enhancing tumor, and (iii) non-enhancing tissue data. These image features are used to create a random forest decision model. The random forest decision model is then used in the prediction to identify tissue types and volumetric calculations.

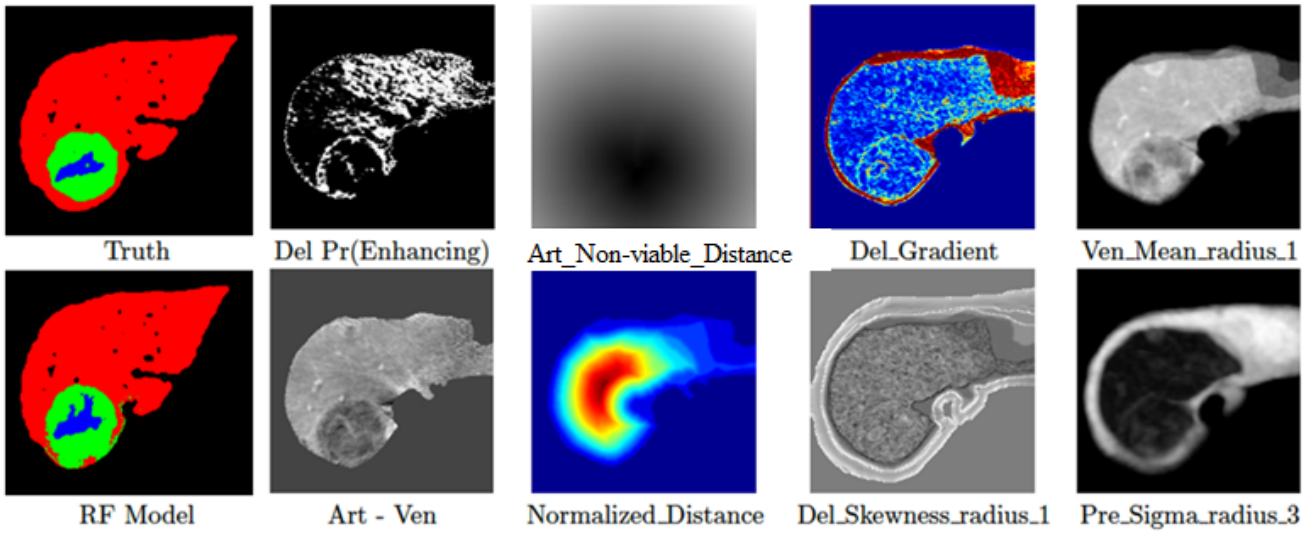
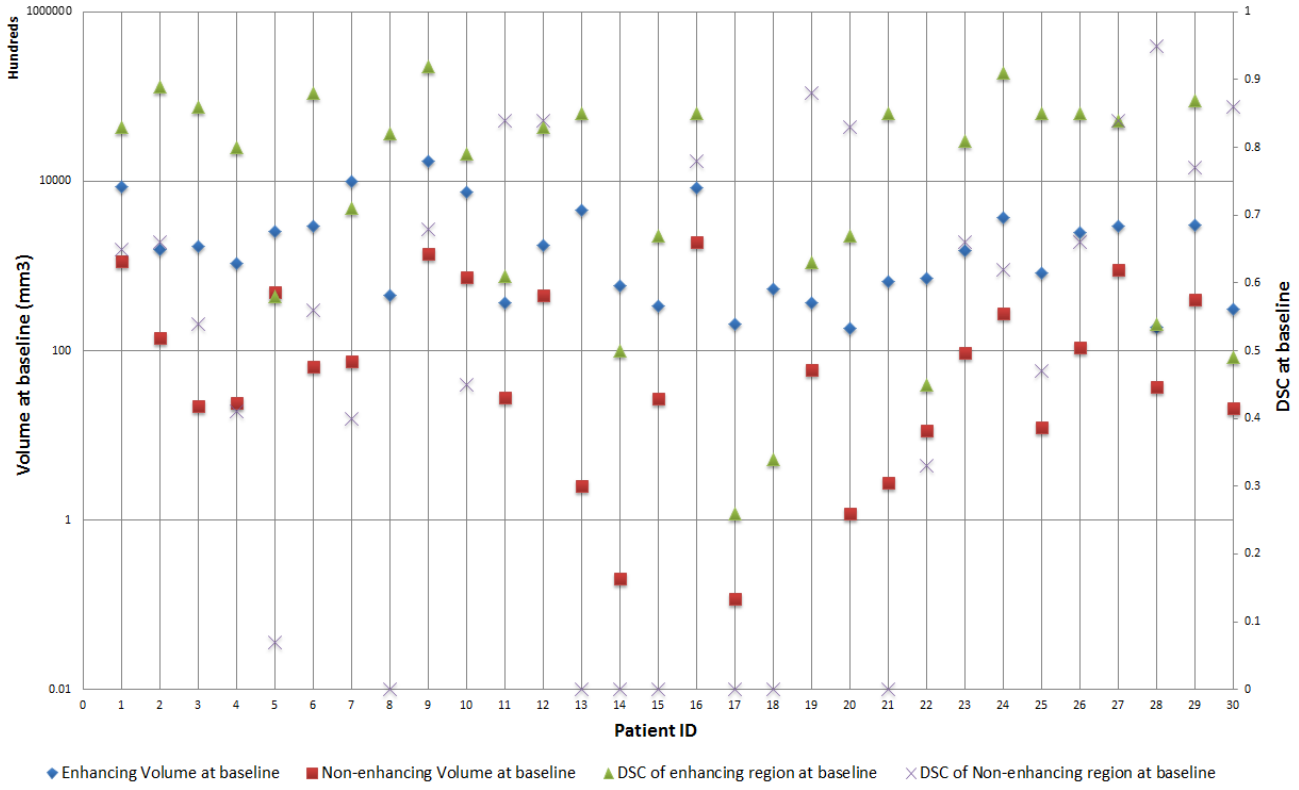


Figure 3: Representative images. Manual segmentation of background liver (red), enhancing tumor (green), and non-enhancing tumor (blue) regions by an experienced radiologist is seen in 'Truth'. The corresponding automated segmentation using random forest model is also provided in 'RF Model'. Representative feature images with the greatest p-values with respect to a Kruskal-Wallis test are also shown. These image features provide the highest discrimination in tissue types with respect to the manually labeled data.


Hundreds
1000000
10000
100
1
0.01
Volume at baseline (mm3)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
DSC at baseline
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Patient ID
Enhancing Volume at baseline
Non-enhancing Volume at baseline
DSC of enhancing region at baseline
DSC of Non-enhancing region at baseline

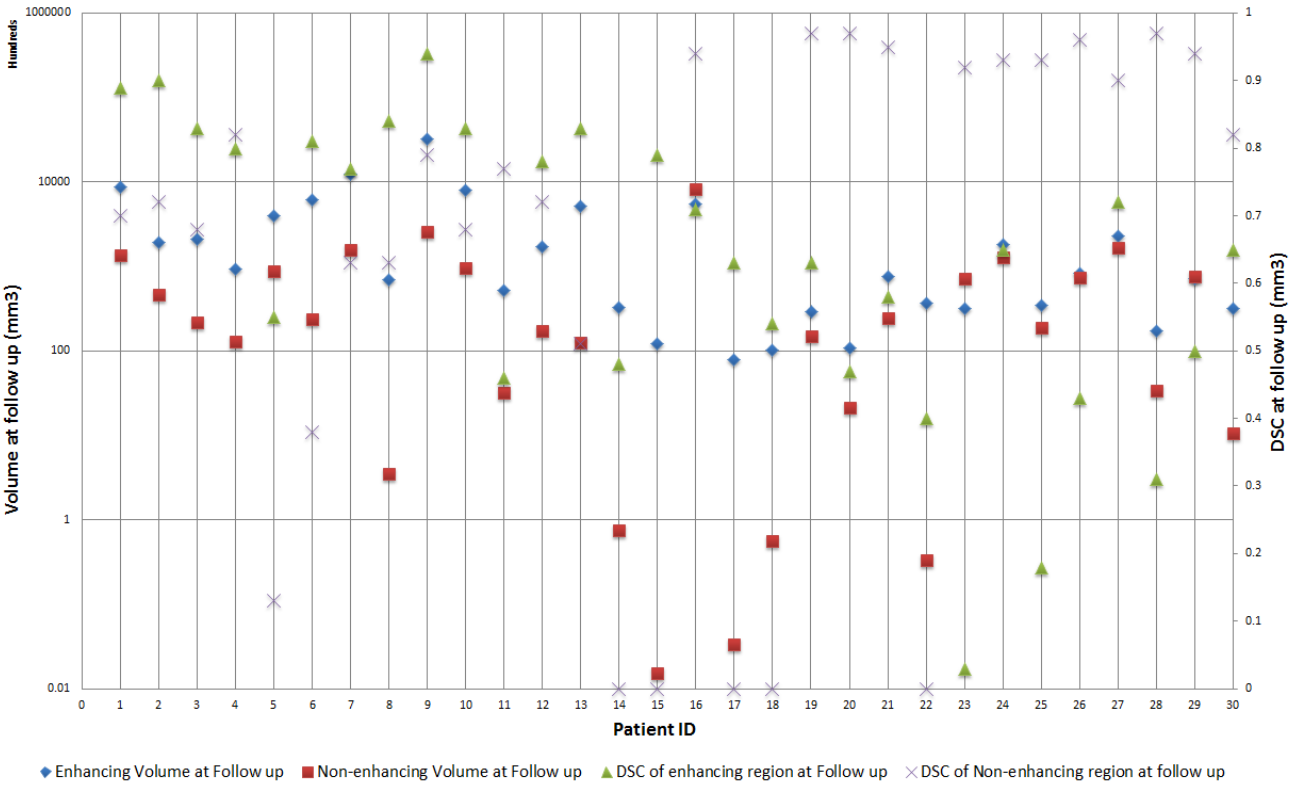

Hundreds
1000000
10000
100
1
0.01
Volume at follow up (mm3)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
DSC at follow up (mm3)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Patient ID
Enhancing Volume at Follow up
Non-enhancing Volume at Follow up
DSC of enhancing region at Follow up
DSC of Non-enhancing region at follow up

Figure 4: A comprehensive overview of the accuracy of the random forest predictions is present. DSC calculations are aligned vertically with the respective patient. DSC overlap is shown between the non-enhancing tissue and enhancing tumor (a) at baseline and (b) at follow up. Median, mean, and standard deviation summary statistics of the dice similarity coefficients are also grouped by treatment modality and shown.

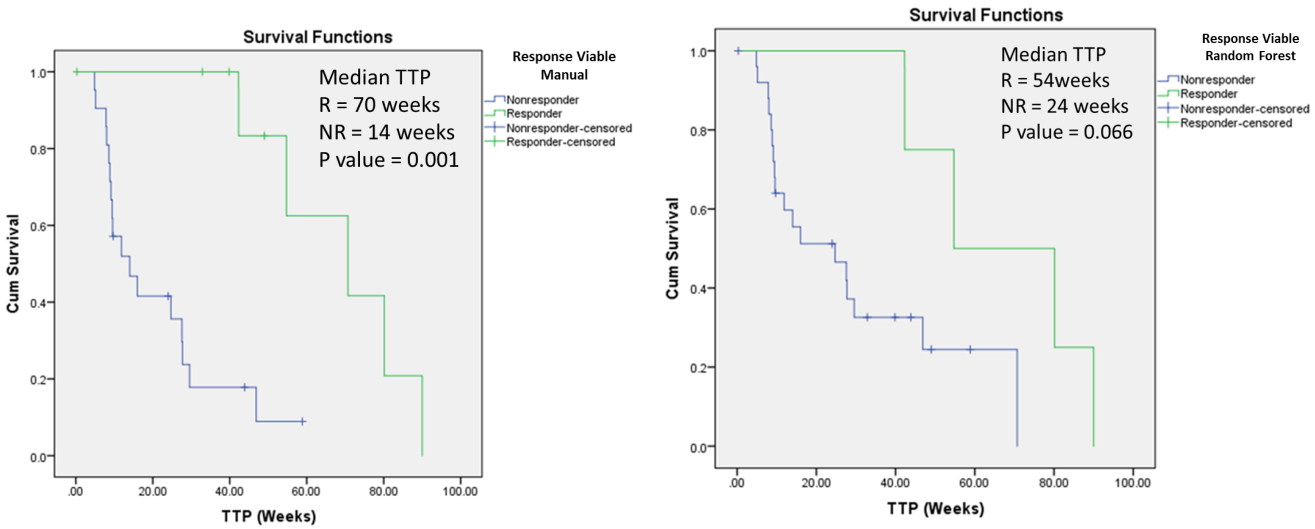


Figure 5: Kaplan-Meier survival curves of responders and nonresponders according to volume change of enhancing tumor estimated by manual segmentation (right) and RF-based automated segmentation (left). Log-rank test showed that both methods have significant ability to classify responders and non-responders with accurate prediction of TTP.

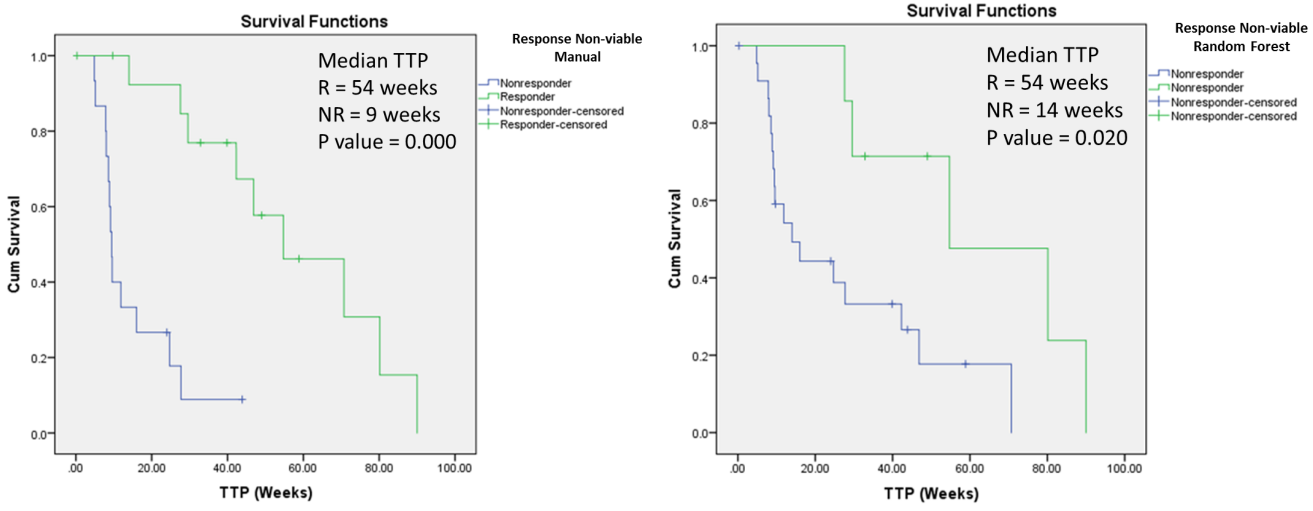


Figure 6: Kaplan-Meier survival curves of responders and nonresponders according to volume change of non-enhancing tumor estimated by manual segmentation (right) and RF-based automated segmentation (left). Log-rank test showed that both methods have significant ability to classify responders and non-responders with accurate prediction of TTP.