Published in final edited form as:
Stat Med. 2020 December 30; 39(30): 4704–4723. doi:10.1002/sim.8749.

Tumor Heterogeneity Estimation for Radiomics in Cancer

Ani Eloyan[a], **Mun Sang Yue**[b], **Davit Khachatryan**[c]

[a]Department of Biostatistics, Brown University School of Public Health, Providence, RI

[b]Department of Biostatistics, Gilead Sciences Inc., Foster City, CA

[c]Division of Mathematics and Science, Babson College, Babson Park, MA

Abstract

Radiomics is an emerging field of medical image analysis research where quantitative measurements are obtained from radiological images that can be utilized to predict patient outcomes and inform treatment decisions. Cancer patients routinely undergo radiological evaluations when images of various modalities including computed tomography, positron emission tomography, and magnetic resonance images are collected for diagnosis and for evaluation of disease progression. Tumor characteristics, often referred to as measures of *tumor heterogeneity*, can be computed using these clinical images and used as predictors of disease progression and patient survival. Several approaches for quantifying tumor heterogeneity have been proposed, including intensity histogram-based measures, shape and volume-based features, and texture analysis. Taking into account the topology of the tumors we propose a statistical framework for estimating tumor heterogeneity using clustering based on Markov Random Field theory. We model the voxel intensities using a Gaussian mixture model using a Gibbs prior to incorporate voxel neighborhood information. We propose a novel approach to choosing the number of mixture components. Subsequently, we show that the proposed procedure outperforms the existing approaches when predicting lung cancer survival.

Keywords

computed tomography; machine learning; cancer imaging; Markov Random Fields; image segmentation

1 Introduction

Quantitative features of medical images are used routinely by clinicians for disease diagnosis, assessment of response to treatment, and for prediction of clinical outcomes such as survival time and disability scores. While in the past a small number of image features (such as tumor size in cancer) were obtained and recorded by radiologists by manually viewing the images, in recent years there is interest in estimating a large number of radiological image features automatically. Aerts et al. (2014) hypothesize that radiological image features obtained from medical imaging data using advanced computational techniques can describe cancerous tissues. The newly emerging field concerned with the estimation of numerous imaging features and subsequent implementation of algorithms to predict clinical outcomes and response to treatment is referred to as *radiomics*.

In cancer imaging, quantitative features of tumors calculated from images of various modalities such as computed tomography (CT), magnetic resonance images (MRI), and positron emission tomography (PET) are often used for diagnosis and prognosis of cancer. Cancer tumors are heterogeneous objects due to their structure, vasculature, cell types within the tumor, and functional characteristics. Herein, the term “tumor heterogeneity” refers to intratumor heterogeneity, i.e. heterogeneity of structures within a single tumor observable on a radiological image. Depending on the imaging modality used in a particular setting, various types of tumor heterogeneity can be observed on the radiological image of a patient. For example, since PET imaging is used to observe metabolic processes in the tumor, features related to tumor metabolism are used as measures of tumor heterogeneity computed from PET scans. On the other hand, since structural MRI captures the structure of the tumor, MRI scans are used to estimate tumor heterogeneity features related to tumor anatomy and shape.

Quantifying tumor heterogeneity is imperative as heterogeneity measures can be used to compare tumors with one another, develop novel targeted therapies, identify longitudinal trends in disease development, determine patient’s response to a particular treatment, and predict clinical outcomes. In addition, medical imaging is performed using noninvasive in-vivo imaging technology, which is routinely performed as part of clinical care. Hence, the analysis of tumor heterogeneity based on medical imaging can potentially be performed using routinely collected images without the need for further data collection.

One common approach to estimating tumor heterogeneity is by using the histogram, in particular the radiomic features based on the histogram of intensities of image voxels [three dimensional (3D) pixel]. A comprehensive review of the literature on histogram measures for quantifying tumor heterogeneity is given by Just (2014). In this approach, a histogram of tumor voxel intensities is calculated and features of the histogram are used as measures of heterogeneity. For example, Figure 1 presents CT scans of patients with lung cancer along with the corresponding intensity histograms. Common features in the literature include skewness, kurtosis, entropy, and percentiles of the tumor intensity histogram. For example, the histogram of tumor intensities of patient 1 in Figure 1 is more left skewed than that of patient 2 with skewness values of −2.47 and −2.13 correspondingly, while the kurtoses of intensities are 6.156 and 3.48 suggesting that the intensities of the tumor of patient 1 have a sharper peak than those of patient 2. Finally, the means (standard deviations) of the two histograms are 961.09 (sd = 201.73) and −49.12 (sd = 234.43) suggesting that the intensities of the first histogram are centered higher than those of the second histogram with similar spread. These mean, standard deviation, skewness, and kurtosis values are then used as measures of tumor heterogeneity to compare tumors. Studies evaluated the relationship of the histogram metrics with response to radiotherapy (Peng et al., 2013), as well as classification of tumor types (Gutierrez et al., 2014). A major shortcoming of these intensity histogram-based approaches is the elimination of information on spatial structure of the tumor. This is potentially a serious limitation as differences in spatial distributions of voxels with similar intensities will be missed by this approach.

We provide a brief (and by no means exhaustive) overview of a few methods for tumor heterogeneity estimation for structural images, data from other imaging modalities, and genomics in this paragraph. Brooks and Grigsby (2013) considered another intensity-based measure of heterogeneity using a pixel-level approach. The proposed measure is based on a distance-dependent mean deviation from a linear intensity gradation. Roberts et al. (2017) proposed the use of wavelet-based scaling indices for prediction of breast cancer diagnosis using mammography imaging. O'Sullivan et al. (2005) discussed the incorporation of tumor shape assessments into estimation of spatial heterogeneity of tumors using fluoro-deoxyglucose-PET data. As discussed above, there are differences in data obtained from PET imaging and mammography and those obtained using the MRI or CT technologies, hence, the methods proposed by Roberts et al. (2017) and O'Sullivan et al. (2005) may not be directly applicable to analyzing tumor heterogeneity using CT imaging as in the motivating dataset in this article. Nevo et al. (2016) discussed issues related to measurement error in biomarker data and approaches for taking into account such errors to reduce misclassification of cancer subtypes. The term "tumor heterogeneity" is widely used in statistical genomics literature to refer to variability in tumor cells. For example, Xu et al. (2015) proposed the MAD Bayes approached for inference on tumor heterogeneity using next-generation sequencing data. Ni et al. (2019) used Bayesian hierarchical varying-sparsity regression models for estimating biomarkers of tumors from proteogenomics.

Aerts et al. (2014) proposed a general approach for using descriptors of tumor heterogeneity for cancer prognosis and tumor classification by building on histogram and shape features. Four types of image features were used to quantify tumor heterogeneity: histogram summaries, shape, texture, and transformation-based features. The prognostic ability of the radiomic features was presented and validated in a large dataset of medical images from patients with lung and head-and-neck cancers. Specifically, cancer survival was considered as an outcome. Textural features were obtained using grey-level co-occurrence and grey-level run-length matrix-based approaches. Once the radiomic features were extracted Aerts et al. (2014) used the Kaplan-Meier product-limit estimator (Kaplan and Meier, 1958) to obtain associations of radiomic features with survival. One radiomic feature from each of the four types of image features was selected based on feature stability analysis to form the radiomic signature consisting of the resulting four representative features. This radiomic signature was then used in a Cox Proportional Hazards model for prediction of survival. Limitations of this approach include the choice of only four features for prediction.

The novel approach for tumor heterogeneity estimation presented in this article builds on the existing estimation methods while accounting for spatial distribution of voxels. On a high level, our proposed approach consists of 2 steps: 1) we propose a new approach for indentification of clusters of voxels with similar intensity profiles (using a methods similar to those for brain tissue segmentation Zhang et al. (2001)), 2) computation of features of the voxel intensities in these estimated clusters as measures of tumor heterogeneity. To model the voxel intensities accounting for their spatial distribution we use Markov Random Fields (MRF) (Li, 2009). We then compute within and between cluster summary measures of voxels with similar intensity profiles that describe tumor heterogeneity. We present the performance of our proposed MRF-based clustering approach in simulation studies. In addition, we show that the proposed tumor heterogeneity estimation approach performs as well as or better than other approaches in ranking images by their heterogeneity using simulated data. We then present the performance of regularized regression when predicting lung cancer patient survival using the novel estimator of tumor heterogeneity and compare the performance of this method to existing approaches using time-dependent receiver operating characteristic (ROC) curves. Specifically, we consider the Cox Proportional Hazards model (Andersen and Gill, 1982) and generalized linear models with Ridge, Lasso, and Elastic-Net penalization. We illustrate that our proposed estimation procedure outperforms other approaches when judged based on maximization of cross-validation area under the curve (AUC) as a criterion of predictive performance of the model. We show that many existing approaches to tumor heterogeneity estimation are special cases of our proposed framework.

The article is organized as follows. Section 2 presents our proposed method, where Section 2.1 provides an overview of the data, Section 2.2 describes notations used throughout the Section. Next, Section 2.3 introduces our proposed novel estimator for tumor heterogeneity, where Section 2.3.1 provides details of our proposed estimation algorithm, Section 2.3.2 shows our proposed procedure for selecting the number of Gaussian mixture components, and Section 2.3.3 presents the summary measures for evaluating tumor heterogeneity. Section 3 presents simulation results on comparison of our proposed methods with other approaches. Section 4 shows the results of the prediction of lung cancer survival using our proposed tumor heterogeneity estimation approach and existing methods. Finally, Section 5 presents concluding remarks.

2 Data and Methods

2.1 Data

In this paper, we examine cancer survival as the clinical outcome of interest. We consider data from patients with non-small cell lung cancer (NSCLC) collected at the MAASTRO Clinic (Maastricht, The Netherlands) and publicly available from The Cancer Imaging Archive (TCIA) located at http://www.cancerimagingarchive.net (Clark et al., 2013). NSCLC is the most common type of lung cancer accounting for about $85 - 90\%$ of lung cancer diagnoses according to the American Cancer Society. The database we obtained from TCIA includes CT scans from 422 patients with inoperable, histologic, or cytologic confirmed NSCLC along with their demographic and clinical information including age, sex, histology of tumors, stage of cancer, and the clinical N and T stages for each participant. The clinical N stage corresponds to the density and spread of lymph nodes and the clinical T stage corresponds to the size and extent of the tumor. Figure 2 presents the distribution of patients in this study by their tumor histology and cancer stage. The average age of the participants in this sample is 68 years with standard deviation of 10.1. There are 290 males and 132 females. A spiral CT with a 3 mm slice thickness is performed for each participant including the complete thoracic region. Further details on data collection parameters and scanning information can be found at Aerts et al. (2015) and hence are not repeated here. Our aim is to develop a tool for prediction of cancer survival, using a novel tumor heterogeneity measure that we define herein.

A few details on preprocessing of medical imaging data deserve a special note. A radiologist reviewed each CT scan and manually delineated the tumor volume by marking voxels visually identified as locations on the tumor surface. In the example illustrated in Figure 3, this delineation starts by marking four of the vertices on the tumor surface. The delineation then continues until enough surface points are obtained to delineate the full tumor surface (as shown in Figure 3 bottom left). For each participant, the coordinates of these tumor surface vertices are saved in a matrix and provided on TCIA in the RTSTRUCT format along with other scan-specific information such as the thickness of the slices, etc. Since in most cases not all vertices are marked manually, there is a need to interpolate the vertices automatically. We pre-processed these files to obtain all tumor voxels by connecting the vertices using linear interpolation in the R software (R Core Team, 2016). As a result, for each participant we obtain a $V_i \times 3$ matrix of coordinates of all tumor voxels as presented in Figure 3 bottom right, where V_i defines the total number of tumor voxels for subject i. We can now use these coordinate matrices to extract the intensities of tumor voxels for each participant. All patients received standard of care treatments for their corresponding cancer stage, however, treatment information was not available in the database.

2.2 Notations and Existing Approaches for Tumor Heterogeneity Estimation

Let I_{iv} define the intensity of the CT image at voxel $v = 1, \dots, V_i$ for subject $i = 1, \dots, N$, where V_i is the number of tumor voxels and N is the total number of patients. While the number of tumor voxels V_i may vary from subject to subject, for simplicity of notation we drop the subscript i and use $V = V_i$. Let $S = \{1, \dots, V\}$ define the set of all tumor voxels. The coordinates of each voxel v are defined by x_v, y_v, z_v. Tumor heterogeneity for subject i is defined as a vector of functions of voxel intensities $\boldsymbol{H}_i = [H_{i1}(I_{iv}), \dots, H_{in}(I_{iv})]$, where n is the total number of functions, where each function corresponds to a radiomic feature. The estimation of tumor heterogeneity can be based on intensity values at voxel level or descriptive statistics at tumor level. We first briefly review and introduce notation for some of the commonly used approaches to estimate $\boldsymbol{H}_i$. Histogram approaches treat the vector of intensities of a tumor $I_{i1}, \dots, I_{iV}$ as an independent and identically distributed set of observations from a density $f_i(x)$. The standardized moments of $f_i(x)$, where the jth moment is denoted by M_{ij}, for $j = 1, \dots, m$, are estimated and used as approximations of tumor heterogeneity. For example, if a study uses mean, variance, skewness, and kurtosis as in Figure 1 then $\boldsymbol{H}_i = [M_{i1}, M_{i2}, M_{i3}, M_{i4}]$. As tumor shape- and volume-based features are commonly used to describe tumor heterogeneity, we define by $S_{i1}, \dots, S_{is}$, s shape- and volume-based measures. Examples of shape measures include estimates of tumor smoothness and sphericity, while the number of voxels on the tumor surface and the number of voxels within the tumor are examples of volume measures. Finally, let $W_{i1}, \dots, W_{iw}$ denote the collection of features quantifying the texture of the tumor image as defined by Aerts et al. (2014). These measures estimate patterns of spatial distributions of intensities using grey level co-occurrence matrices (Haralick et al., 1973) or wavelet-based approaches (see the Supplementary Materials of Aerts et al. (2014) for more details).

2.3 Neighborhood-Based Tumor Heterogeneity

We propose a novel Neighborhood-Based Tumor Heterogeneity (NBTH) estimation method for measuring tumor heterogeneity. This approach is motivated by the intuition that: 1) a tumor is a collection of k locally homogeneous regions within the image labeled as $\Lambda = \{1, \dots, k\}$, where k may be known or estimated; 2) the descriptive characteristics of these locally homogeneous regions are related to cancer outcomes such as survival. Note, a "region" within a tumor can be thought of as a collection (or cluster) of voxels. In addition, the term "locally homogeneous region" indicates that the voxels within this region have similar intensities. After estimating locally homogeneous regions within the tumor, i.e. estimating the cluster membership label $\lambda_V \in \Lambda$ for each voxel $v \in S$, we use summary characteristics of those regions to define our novel measure of tumor heterogeneity. Hence, identifying locally homogeneous clusters of voxels with similar intensity profiles and computing the characteristics of the resulting clusters is used to measure tumor heterogeneity.

In image analysis, it is common to assume that voxels in a close vicinity of each other are likely to have similar intensities. MRFs are undirected graphical models widely used in image analysis (i.e. for segmentation of different tissues in the brain Zhang et al. (2001), Wang (2012), Zhang and Ji (2009) in computer vision problems), as they provide a modeling framework for correlated entities in space such as the intensities of voxels in images. Let $\mathscr{N} = \mathscr{N}(v)$, $v \in S$ define a neighborhood system, where $\mathscr{N}(v)$ is a neighborhood of a tumor voxel v such that $v \notin \mathscr{N}(v)$ and $v \in \mathscr{N}(v_1) \Leftrightarrow v_1 \in \mathscr{N}(v)$, i.e. voxel v does not belong in its neighborhood and for any other voxel v_1, v belongs in the neighborhood of v_1 if and only if v_1 belongs in the neighborhood of v. Each voxel $v \in S$ is assigned a true and unknown label $\lambda_V \in \Lambda$ that determines the cluster membership of voxel v. The goal is to estimate the unknown labels λ_V. Let $P(\lambda)$ denote a probability measure on the set of all possible labels Λ. We assume that the labeling field Λ is an MRF on S with respect to the neighborhood system $\mathscr{N}$, in other words

$$\forall \lambda \in \Lambda, P(\lambda) > 0 \text{ and } P\left(\lambda_v \mid \bar{\lambda}_{S \backslash \{v\}}\right) = P\left(\lambda_v \mid \bar{\lambda}_{N(v)}\right).$$

where $\bar{\lambda}_{S \backslash \{v\}}$ denotes the set of cluster labels of all voxels in S excluding v and $\bar{\lambda}_{N(v)}$ denotes the set of cluster labels of all voxels in the neighborhood of v defined as $N(v)$. In other words, we assume that the probability of assigning the label λ_V to voxel v depends on the labeling of the voxels in its neighborhood $\mathscr{N}$ only. A probabilistic model commonly implemented in image segmentation problems is the Finite Gaussian Mixture Model (GMM) Wang (2012), Zhu et al. (2003). FGMM models assume that the voxel intensities are independent and Gaussian mixture densities are used to model the distribution of voxel intensities. This is a convenient model to implement in our context as the voxel labels can be easily estimated, e.g. by using the expectation-maximization (EM) algorithm (Dempster et al., 1977).

In what follows, we will describe a GMM-based MRF modeling framework, and show how the cluster labels will be determined as a result. We will also introduce a method for finding an appropriate number of mixture components in GMM. The clusters emerging from this MRF-GMM modeling will be subsequently used in the definition of the novel measure of tumor heterogeneity.

2.3.1 Estimation of Mixture Component Weights in MRF—In this section, we propose a novel approach for modeling the voxel labels using the GMM framework while incorporating neighborhood voxel intensity information within the MRF framework.

Without loss of generality and for clarity of notation, we drop the index for study subjects denoted herein by i. In GMM, the density function at the observation I_V at voxel $v = 1, \ldots, V$ is modeled using a Gaussian mixture distribution as follows.

$$f(I_v \mid \Pi, \mu_K, \sigma_K) = \sum_{j=1}^{K} \pi_{j,v} \frac{1}{\sigma_K} \phi\left(\frac{I_v - \mu_{j,K}}{\sigma_K}\right) \quad (1)$$

where $\phi(\cdot)$ is the probability density function of a normally distributed random variable with mean 0 and variance 1, $\boldsymbol{\mu}_K = (\mu_{1,K}, \ldots, \mu_{K,K})^T$ is the vector of the mixture component means, and σ_K is the common standard deviation. In this section, we assume that the number of mixture components, defined by K, is known. Finally, $\pi_{j,v}$, $j = 1, \ldots, K$ and $v = 1, \ldots, V$ is the probability that voxel v has label j and Π is the $V \times K$ matrix of voxel specific probabilities $\pi_{j,v}$, such that $\sum_{j=1}^{K} \pi_{j,v} = 1$, for all $v \in 1 ..., V$. Let Z_V denote the latent label of vth voxel, then $P(Z_V = j) = \pi_{j,V}$. Following the approach proposed by Eloyan and Ghosh (2011) and extended to high dimensional settings by Meng and Eloyan (2017) for estimation of Gaussian mixture based approximations of density functions, we assume the means and variance of the mixture elements defined by $\mu_{1,K}, \ldots, \mu_{K,K}$ and σ_K are fixed. Eloyan and Ghosh (2011) propose fixing the values of $\mu_{1,K}, \ldots, \mu_{K,K}$ at equidistant points on the support of the empirical distribution of the observed data points, while Meng and Eloyan (2017) propose using the k-means algorithm for selecting the means of the Gaussian mixture components using the observed data. The mixture standard deviation σ_K is set as a small value relative to the differences between the mixture element means $\boldsymbol{\mu}_K$. This approach results in a computationally efficient estimation of the mixture weights and the resulting density estimate approximates the true underlying density of the observed data in terms of minimizing the Kullback-Leibler Divergence (KLD) between the true and estimated densities. In addition, as shown by Eloyan and Ghosh (2011) and Meng and Eloyan (2017) a hypothesis testing procedure can be implemented to estimate the number of mixture components, K, within this framework. Both of these models assume that the observed voxel intensities are statistically independent. This assumption may not hold in image analysis as intensities of neighboring voxels are often close to each other due to smoothness of the image. To take into account the spatial information using the Markov random field theory, we extend the model and incorporate the spatial smoothness of the image by modeling the prior distribution of the probabilities $\pi_{j,V}$ for individual voxels using a Gibbs function as follows.

$$p(\Pi) = \frac{1}{T} e^{-U(\Pi)}$$

where $U(\Pi) = \beta \sum_{v=1}^{V} V_{\mathcal{N}(v)}(\Pi)$ is the smoothing prior with regularization parameter β, T is the normalizing constant of the density function, $V_{\mathcal{N}(v)}(\Pi)$ is the clique potential function (Li, 2009). Blekas et al. (2005) and Nguyen and Wu (2012) discuss various forms of the smoothing prior $U(\Pi)$ and the effect of the choice of $U(\Pi)$ on parameter estimation. In this paper, we implement the smoothing prior proposed by Nguyen and Wu (2012) as it provides a computationally feasible approach for parameter estimation in our setting, where tumors are observed in 3D space and the number of voxels in some tumors observed in the NSCLC dataset analysed in this manuscript is larger than 80,000. The proposed algorithm is based on an implementation of the EM-algorithm to obtain *a posteriory* MAP estimates of the parameters of interest. In the kth step of the iterative EM-algorithm estimation procedure we implement the following two steps.

E-step: We compute the expectation of the log-likelihood given the observed values and the values of the coefficients from the previous iteration. We define $f_{j,K}(x) = \phi\left(\frac{I_v - \mu_{j,K}}{\sigma_K}\right)$ and $\boldsymbol{I} = (I_1, \ldots, I_V)^T$. By Bayes rule we obtain.

$$P\left(Z_v = j \mid x, \Pi^{(k)}\right) = \frac{\pi_{j,v}^{(k)} f_{j,K}(I_v)}{\sum_{j=1}^{K} \pi_{j,v}^{(k)} f_{j,K}(I_v)} = w_{j,v}\left(\Pi^{(k)}, I\right).$$

The expectation of the log-likelihood of the complete data can be derived as follows.

$$Q\left(\Pi \mid \Pi^{(k)}\right) = \sum_{v=1}^{V} \sum_{j=1}^{K} w_{j,v}\left(\Pi^{(k)}, I\right)\left\{\log\left[f_{j,K}(I_v)\right] + \log\left(\pi_{j,v}\right)\right\} + \log[p(\Pi)]. \qquad (2)$$

The smoothing prior (Nguyen and Wu, 2012) is defined as follows.

$$U(\Pi) = -\sum_{v=1}^{V} \sum_{j=1}^{K} G_{j,v}^{(k)} \log \pi_{j,v} \text{ where } G_{j,v}^{(k)} = exp\left(\frac{\beta}{2|\mathcal{N}(v)|} \sum_{v' \in \mathcal{N}(v)} \left(w_{j,v'}\left(\Pi^{(k)}, I\right) + \pi_{j,v'}^{(k)}\right)\right)$$

where $|\mathcal{N}(v)|$ denotes the number of elements in the neighborhood set $\mathcal{N}(v)$.

M-step: We maximize the function Q in (2) under constraints $\sum_{j=1}^{K} \pi_{j,v} = 1$. To obtain probabilities that satisfy these conditions Blekas et al. (2005) propose a projection of the estimated probabilities to the corresponding space. In our implementation we use the Lagrange Multiplier method to incorporate these constraints directly in the M-step of the EM-algorithm. Using the constants λ_V as coefficients for the constraints on probabilities we build the Lagrangian function to be maximized as follows.

$$Q_\lambda\left(\Pi, \Pi^{(k)}\right) = Q\left(\Pi \mid \Pi^{(k)}\right) + \sum_{v=1}^{V} \lambda_v \left(1 - \sum_{j=1}^{K} \pi_{j,v}\right).$$

By taking the derivatives with respect to $\boldsymbol{\pi}_{j,V}$ for $j = 1, \ldots, K$, and λ_V we obtain the following system of equations.

$$\begin{aligned}\frac{\partial Q_\lambda\left(\Pi, \Pi^{(k)}\right)}{\partial \pi_{j,v}} &= \frac{w_{j,v}\left(\Pi^{(k)}, I\right)}{\pi_{j,v}} + \frac{G_{j,v}^{(k)}}{\pi_{j,v}} - \lambda_v = 0 \\ \frac{\partial Q_\lambda\left(\Pi, \Pi^{(k)}\right)}{\partial \lambda_v} &= 1 - \sum_{j=1}^{K} \pi_{j,v} = 0\end{aligned} \tag{3}$$

We first derive $\boldsymbol{\pi}_{j,V}$ from equation (3) and substitute the expression in the next V equations to obtain $\hat{\lambda}_v = 1 + \sum_{j=1}^{K} G_{j,v}^{(k)}$. Then by substituting these values in the expression for $\hat{\pi}_{j,v}^{(k+1)}$ we obtain.

$$\hat{\pi}_{j,v}^{(k+1)} = \frac{w_{j,v}\left(\Pi^{(k)}, I\right) + G_{j,v}^{(k)}}{\hat{\lambda}_v}. \tag{4}$$

The algorithm stops when the absolute norm difference of the updated probability matrix $\widehat{\Pi}^{(k+1)}$ and that computed in the previous iteration $\widehat{\Pi}^{(k)}$ is smaller than a predefined small value $\epsilon > 0$. Finally, if the algorithm stops at iteration $k+1$, cluster labels $\hat{\lambda}_1, \ldots, \hat{\lambda}_V$ are assigned based on the estimated probabilities as follows $\hat{\lambda}_v = j$, if $\hat{\pi}_{j,v}^{(k+1)}$ is highest among $\hat{\pi}_{1,v}^{(k+1)}, \cdots, \hat{\pi}_{K,v}^{(k+1)}$. The above determination of labels results in clusters that account for the spatial distribution of voxel intensities. Our novel measure of tumor heterogeneity is directly based on statistics of those clusters. An important question that still remains is the choice of the number of mixture components defined by K assumed to be fixed in the beginning of this Section. In the subsection that follows, we relax that assumption and propose an iterative procedure for selecting K given a starting value of K_0 based on our goal of minimizing the distance between the true density of intensities $f(\boldsymbol{I})$ and the estimated density using our proposed approach.

2.3.2 Selection of the Number of Mixture Components—

We implement a hypothesis testing procedure for estimation of the number of Gaussian mixture components in the proposed MRF-GMM procedure with the goal to minimize the distance between the true probability density function $f(I_V)$ of the image intensities and the estimated function defined as $\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)$. The sample estimate of KLD between true and estimated functions is $\frac{1}{V}\sum_{v=1}^{V} \log \frac{f(I_v)}{\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)}$. The intuition for our proposed procedure is that this sample estimate cannot be directly minimized in K since the true underlying density is unknown. However, we can compute the difference of sample estimates of KLD between true and estimated density functions of two consecutive values of K, i.e. $\frac{1}{V}\sum_{v=1}^{V} \log \frac{f(I_v)}{\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)} - \frac{1}{V}\sum_{v=1}^{V} \log \frac{f(I_v)}{\hat{f}_{K+1}\left(I_v \mid \widehat{\Pi}, \mu_{K+1}, \sigma_{K+1}\right)}$. Hence, to estimate the

$= \frac{1}{V}\sum_{v=1}^{V} \log \frac{\hat{f}_{K+1}\left(I_v \mid \widehat{\Pi}, \mu_{K+1}, \sigma_{K+1}\right)}{\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)}$

number of components in the mixture defined by K, we consider an iterative approach where, starting from a predefined value K_0, we compute the estimated function

$\hat{f}_{K0}\left(I_v \mid \widehat{\Pi}, \mu_{K0}, \sigma_{K0}\right)$ using the steps of the EM-algorithm described in the previous subsection iteratively until convergence. Further, at each iteration K of the procedure we increase the number of components in the mixture by 1, estimate $\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)$, and test whether the expected value of the difference of KLD of estimated function at iteration K and true underlying probability function and KLD of estimated function at iteration $K + 1$ and true probability function is equal to zero as follows.

$$H_0: E\left[\log\frac{\hat{f}_{K+1}\left(I \mid \widehat{\Pi}, \mu_{K+1}, \sigma_{K+1}\right)}{\hat{f}_K\left(I \mid \widehat{\Pi}, \mu_K, \sigma_K\right)}\right] = 0, \tag{5}$$

for each $K \geq K_0$, where K_0 is a starting value of the number of components. As K increases by 1, we define the empirical estimate of the difference between KLD of $\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)$ from the true function f and that of $\hat{f}_{K+1}\left(I_v \mid \widehat{\Pi}, \mu_{K+1}, \sigma_{K+1}\right)$ as follows for voxel v.

$$\widehat{\Delta}_{K,v} = \log\frac{\hat{f}_{K+1}\left(I_v \mid \widehat{\Pi}, \mu_{K+1}, \sigma_{K+1}\right)}{\hat{f}_K\left(I_v \mid \widehat{\Pi}, \mu_K, \sigma_K\right)} \tag{6}$$

for $v = 1, \ldots, V$. Under regularity conditions (presented by Eloyan and Ghosh (2011)) we can use the following rule for testing the hypothesis in (5) and choosing K.

$$Z_{K,v} = \sqrt{V}\frac{\overline{\Delta}_{K,v}}{S_{\Delta K,v}} \tag{7}$$

where $\overline{\Delta}_{K,v}$ is the sample mean and $S_{\Delta K,V}$ is the sample variance of $\widehat{\Delta}_{K,v}$. We reject the null hypothesis in (5) if $z_{K,V} > z_\alpha$, where z_α denotes the α% upper percentile of the standard normal density for a pre-specified value of α. The procedure stops when we fail to reject the the null hypothesis or if the number of cluster labels is greater than a pre-specified maximum value of number of mixture components defined as K_m.

2.3.3 Summary Measures

2.3.3 Summary Measures—After having outlined the process of voxel clustering that accounts for the spatial distribution of intensities, we are in a position to define our novel measure of tumor heterogeneity using various summary statistics emerging from those clusters. Our proposed estimator of tumor heterogeneity consists of three sets of summary measures derived from the clustering results. One set of measures relates to the mean and variability of cluster sizes, the second set relates to the summarized intensities, while the third set of features relates to cluster shapes. For example, the variability of estimated cluster sizes for a tumor with highly variable-sized clusters is expected to be large, while the variability of cluster sizes of a tumor with similarly sized intensity clusters is expected to be small. In addition to more commonly used summary measures to quantify tumor heterogeneity described below, we use the Rao's diversity index to measure the diversity of intensities in each cluster. Rao's Quadratic Entropy (Rao, 1982) is a statistical measure defined as the expected difference between two objects randomly selected from a given population. The difference between the objects can be quantified using any nonnegative function, including distance functions. The index will be used herein to quantify the diversity of voxel intensities within clusters of a tumor.

Let the vector $[K_{i1}, \ldots, K_{iK}]^T$ define the cluster sizes, the vectors $\left[I_{i1}^{mean}, \ldots, I_{iK}^{mean}\right]^T$ and $\left[I_{i1}^{sd}, \ldots, I_{iK}^{sd}\right]^T$ define the means and standard deviations of intensities within each cluster, the vector $\left[I_{i1}^{Rao}, \ldots, I_{iK}^{Rao}\right]^T$ define the Rao's index of the intensities in each cluster, and $\left[I_{i1}^{ent}, \ldots, I_{ik}^{ent}\right]^T$ define the within cluster entropy of intensities, i.e.

$K_{i1} = \sum_{v=1}^{V} \mathrm{I}\left[\hat{\lambda}_{iv} = 1\right], \ldots, K_{iK} = \sum_{v=1}^{V} \mathrm{I}\left[\hat{\lambda}_{iv} = K\right],$

$I_{i1}^{mean} = \frac{1}{K_{i1}} \sum_{v=1}^{V} I_{i,v} \cdot \mathrm{I}\left[\hat{\lambda}_{iv} = 1\right], \ldots, I_{iK}^{mean} = \frac{1}{K_{iK}} \sum_{v=1}^{V} I_{iv} \cdot \mathrm{I}\left[\hat{\lambda}_{iv} = K\right],$

$I_{i1}^{sd} = \sqrt{\frac{1}{K_{i1}-1} \sum_{v=1}^{V} \left(I_{i,v} - I_{i1}^{mean}\right)^2 \cdot \mathrm{I}\left[\hat{\lambda}_{iv} = 1\right]}, \ldots, I_{iK}^{sd},$

$= \sqrt{\frac{1}{K_{iK}-1} \sum_{v=1}^{V} \left(I_{i,v} - I_{i1}^{mean}\right)^2 \cdot \mathrm{I}\left[\hat{\lambda}_{iv} = K\right]}$

$I_{i1}^{Rao} = \sum_{v=1}^{V} \sum_{v'=1}^{V} P_1\left(I_{i,v}\right) D_{i,v,v'} P_1\left(I_{i,v'}\right) \mathrm{I}\left[\hat{\lambda}_{iv} = 1\right] \mathrm{I}\left[\hat{\lambda}_{iv'}=1\right], \ldots,$ and
$I_{iK}^{Rao}=\sum_{v=1}^{V} \sum_{v'=1}^{V} P_K\left(I_{i,v}\right) D_{i,v,v'} \mathrm{P}_{\mathrm{K}}\left(I_{I,v'}\right) \mathrm{I}\left[\hat{\lambda}_{iv}=\mathrm{K}\right] \mathrm{I}\left[\hat{\lambda}_{iv'}=K\right]$

$I_{i1}^{ent} = \sum_{v=1}^{V} P_1\left(I_{i,v}\right) \log P_1\left(I_{i,v}\right) \cdot \mathrm{I}\left[\hat{\lambda}_{iv} = 1\right], \ldots, I_{iK}^{ent} = \sum_{v=1}^{V} P_K\left(I_{iv}\right) \log P_K\left(I_{i,v}\right) \cdot \mathrm{I}\left[\hat{\lambda}_{iv} = K\right],$
where $P_j(\cdot)$ is the empirical distribution of the observed intensities in cluster j and $D_{i,v,v'}$ denotes the Euclidean distance between the intensities $I_{i,V}$ and $D_{i,v,v'}$. For a given total number of clusters K, to obtain the first set of summary measures related to cluster size we compute the following measures.

$$C_i^{mean} = mean(K_{i1}, \ldots, K_{ik}) \text{ and } C_i^{sd} = sd(K_{i1}, \ldots, K_{ik}),$$

where C_i^{mean} measures the average cluster size and C_i^{sd} measures the variation among the cluster sizes. We define the second set of measures related to average intensities of the clusters as follows.

$$I_i^{mean1} = mean\left(I_{i1}^{mean}, \ldots, I_{ik}^{mean}\right) \text{ and } I_i^{mean2} = sd\left(I_{i1}^{mean}, \ldots, I_{ik}^{mean}\right),$$

$$I_i^{sd1} = mean\left(I_{i1}^{sd}, \ldots, I_{ik}^{sd}\right) \text{ and } I_i^{sd2} = sd\left(I_{i1}^{sd}, \ldots, I_{ik}^{sd}\right),$$

$$I_i^{Rao1} = mean\left(I_{i1}^{Rao}, \ldots, I_{ik}^{Rao}\right) \text{ and } I_i^{Rao2} = sd\left(I_{i1}^{Rao}, \ldots, I_{ik}^{Rao}\right),$$

$$I_i^{ent1} = mean\left(I_{i1}^{ent}, \ldots, I_{ik}^{ent}\right) \text{ and } I_i^{ent2} = sd\left(I_{i1}^{ent}, \ldots, I_{ik}^{ent}\right),$$

Finally, the third set of summary measures that will complete our definition of tumor heterogeneity is defined as $S_{i1}, \ldots, S_{i8}$. These correspond to features commonly used to describe tumor shape and volume computed for each cluster and averaged across the tumor. The formulae for computing the shape features are described in (Aerts et al., 2014) and repeated here for completeness. The tumor volume is defined as S_1, and computed by counting the number of tumor voxels and multiplying by the volume of each voxel. The tumor surface area S_2 is computed by triangulation. The rest of the shape and size features are functions of S_1 and S_2. Surface to volume ratio $S_3 = \frac{S_2}{S_1}$. Sphericity of the tumor $S_4 = \frac{\pi^{1/3}(6S_1)^{2/3}}{S_2}$. Spherical disproportion $S_5 = \frac{S_2}{4\pi R^2}$, where R is the radius of a tumor sized sphere. Maximum 3D diameter of the tumor is defined as S_6. Compactness $S_7 = 36\pi\frac{S_1^2}{S_2^3}$ and $S_8 = \frac{S_1}{\sqrt{\pi}S_2^{2/3}}$.

As a result of our proposed procedure, we obtain the following vector as an estimator of tumor heterogeneity, for subject $i = 1, 2, \ldots, N$.

$$\widehat{H}_{i,NBTH} = \left(C_i^{mean}, C_i^{sd}, I_i^{mean1}, I_i^{mean2}, I_i^{sd1}, I_i^{sd2}, I_i^{Rao1}, I_i^{Rao2}, I_i^{ent1}, I_i^{ent2}, S_{i1} \ldots, S_{i8}\right)^T \in R^{18 \times 1}$$

When implementing the proposed algorithm, we choose the initial set of means of for the EM-algorithm described in Section 2.3.1 using the k-means algorithm Friedman et al. (2001). We select the maximum number of means defined by K_m by choosing the value of K corresponding to 80% of variance explained in the k-means procedure. Starting from a given initial value of $K_0 = 2$, we use our proposed selection procedure described in Section 2.3.2 to find the number of mixture components. Our proposed procedure for computing $\widehat{H}_{i,NBTH}$ is computationally efficient in relatively low dimensions and computationally feasible in high-dimensional settings. It took about 5 minutes to compute $\widehat{H}_{i,NBTH}$ for a subject with a large tumor (over 80,000 tumor voxels) in our dataset using the R software on a 1.7GHz Dual-Core Intel Core i7 processor. Next, we show simulation studies to illustrate the superior performance of the proposed method as compared with Gaussian Mixture Models and k-means clustering.

3 Simulations

We evaluate the performance of our proposed algorithm in two simulation studies: 1) we evaluate the accuracy of the proposed MRF-GMM algorithm in terms of estimation of voxel labels in an image, i.e. image segmentation, 2) we evaluate the performance of our proposed NBTH algorithm for the estimation of tumor heterogeneity in terms of ranking two images by their heterogeneity. In the first simulation study we compare the performance of our proposed MRF-GMM procedure in terms of its accuracy of labeling voxels to two existing methods - k-means and GMM (Friedman et al., 2001). K-means clustering (implemented using the kmeans function in the R software), is a distance based approach which, among numerous other applications, is used in image segmentation Dhanachandra et al. (2015) for identifying voxel labels to avoid the assumption of Gaussianity of image voxel intensities in the GMM model. Let $\boldsymbol{c} = (c_1, \ldots, c_K)$ denote the vector of cluster centers for a given K indicating the number of clusters for the subject's image. The k-means procedure is based on the intuition that a cluster should contain voxels with similar intensity profiles. The loss function $L(c) = \frac{1}{V}\sum_{v=1}^{V} \min_{1 \le j \le K} \|I_v - c_j\|^2$ can be considered for minimization, and based on $L(\boldsymbol{c})$ the algorithm identifies the set of cluster centers $\hat{c}_1, \ldots, \hat{c}_K$. The voxel labels are then assigned based on the estimated centers. We select the number of clusters in the k-means algorithm by choosing the value of K corresponding to 80% of variance explained. The GMM approach (implemented using the ClusterR package in the R software) is based on the model given in (1), while the voxel intensities are assumed to be statistically independent. We select the optimal number of clusters when implementing GMM using the Akaike information criterion (AIC), specifically, the optimal number of clusters, K, is the value of number of clusters such that an increase in K by 1 results in a change in AIC of less than 1%.

To compare the three methods in terms of clustering accuracy, we use two images used to evaluate image segmentation algorithms by Blekas et al. (2005) and Nguyen and Wu (2012) and reproduced in Figure 4. The goal of the analysis is segmentation of the three clusters in true image 1 and the four clusters in true image 2, where a cluster consists of intensities of the same color in the corresponding image. In the context of tumor imaging, the cluster would correspond to a specific tissue consisting of voxels with similar intensities. For each image, we run 100 simulations where in each simulation we add random noise to the images. To investigate the effect of the magnitude of noise on the performance of the three competitor methods we include two settings of noise for each image as follows: we add normally distributed random noise with standard deviation of $\boldsymbol{\sigma} = 1$ in the first setting and standard deviation of $\boldsymbol{\sigma} = 2$ in the second setting to true image 1 and normally distributed random noise with standard deviation of $\boldsymbol{\sigma} = 2$ in the first setting and standard deviation of $\boldsymbol{\sigma} = 3$ in the second setting to true image 2. After adding the random noise, in order to obtain relatively smooth images resembling real images, we apply a blur to the resulting noisy images where the blurring kernel is the isotropic Gaussian kernel with standard deviation of 1.

For each simulation study, after estimating the cluster labels with each of the three methods, GMM, k-means, and our proposed MRF-GMM approach, we compute four measures to evaluate the accuracy of the estimated image, compared to the true image: 1) cluster overlap corresponding to the proportion of overlapping cluster voxels between the true and estimated images, 2) Jaccard Index that measures similarity between two images by computing the intersection of the voxel sets and dividing by their union, 3) Mutual Information (MI) between the clustering partitions, i.e. let $\widehat{\Lambda}_M$ denote the set of cluster labels obtained using our proposed procedure and Λ_T denote that of the true image then the MI between the two sets is computed as $\sum_{j=1}^{K}\sum_{l=1}^{K} p(j,l)\log\frac{p(j,l)}{p_M(j)p_T(l)}$, where $p(j, l)$ is the probability that the label of a point is in both the cluster labeled by j in $\widehat{\Lambda}_M$ and the cluster labeled by l in Λ_T,

$p_M(j)$ is the probability that the label of a point belongs to the cluster indexed by j in $\widehat{\Lambda}_M$, $p_T(l)$ is the probably that the label of a point belongs to the cluster indexed by l in Λ_T, and 4) Adjusted Rand index (Wagner and Wagner, 2007). Higher values indicate better performance of the algorithm in terms of labeling accuracy. The mean and standard deviation of each of the clustering comparison measures are presented in Table 1 while the boxplots are shown in Figure 4. As shown by all approaches, our proposed MRF-GMM method outperforms the competitors significantly in clustering performance except for image 2 with low level added noise. Nguyen and Wu (2012) recommends setting the smoothing regularization parameter β to 12. We selected values of β at 4, 6, and 12 in our simulations and did not find a significant difference in the results in most cases.

Our second simulation study evaluates the performance of our proposed NBTH method in terms of ranking images according to tumor heterogeneity. The studies based on histogram measures commonly used in the literature postulate that a lower value of uniformity of the intensities, a higher entropy, higher standard deviation of intensities, higher kurtosis values, and positive skewness all represent increased heterogeneity and are related to poorer prognosis (Davnall et al., 2012). When comparing the heterogeneity of two images using our proposed NBTH method we rank the two images by using each of the 18 $\widehat{H}_{i,NBTH}$ features. The final ranking of the images is assigned based on the ranking of the majority of NBTH features. Figure 5 presents the images we used for comparisons for the following two cases. We use very simplistic images in our simulations to ensure that the ranking of images by their heterogeneity is clearly visible.

Case 1: we generate four 50 × 50 matrices shown in Figure 5 row (a) as follows: A_1 is a matrix of zeros (column 1), A_2 is a matrix of zeros, with a diamond of 1s containing 145 pixels (column 2), A_3 is a matrix of 0s with a diamond of 1s and additional smaller overlaid layers of 2s and 3s, A_4 is a matrix of 0s with a diamond of 1s and additional layers up to 6. In summary, the images are generated such that they are ordered in terms of their heterogeneity, i.e. A_4 is more heterogeneous than the rest, A_3 is more heterogeneous than A_2 and A_1, finally, A_2 is more heterogeneous than A_1. We run 500 simulations. In each simulation, we add random Gaussian noise with mean 0 and variance 1 to the images. We compare our proposed method to histogram measures including skewness, kurtosis, and entropy in terms of ranking these images in their heterogeneity. We found that all four methods (skewness, kurtosis, entrophy, and NBTH) correctly ranked images when comparing A_4 to A_1 and when comparing A_3 to A_1 as having higher heterogeneity than those based on A_1 in 100% of simulations. When comparing A_2 based images to those based on A_1 skewness and kurtosis ranked them correctly in terms of heterogeneity in 99% of simulations, entropy in 89%, and our proposed NBTH method in 95% simulations. In summary, our proposed method performs as well as the competitor methods in terms of ordering images in terms of heterogeneity in this simulation study.

Case 2: we use 2 images to compare the effect of higher noise on the estimation of tumor heterogeneity shown in rows (b) and (c) in Figure 5. We again run 500 simulations. The true images are shown in column 1, the images are identical except for the additional white colored collection of voxels in the image in row (c) of column 1 referred to as image 2 which implies that the image in row (c) is more heterogeneous than that in row (b) referred to as image 1. When the added random Gaussian noise has standard deviation of 1 (column 2) skewness correctly classifies image 2 as having higher heterogeneity than image 1 in 3.6% of simulations, kurtosis in 8.2%, entropy in 17% and our proposed method in 77.4%. Further, when the added noise has standard deviation 1.5 (column 3) skewness correctly ranks the images by their heterogeneity in 22% of simulations, kurtosis in 48.4%, entropy in 34.8%, and our proposed method in 89% of simulations. Finally, when the added noise has standard deviation 2 (column 4) then skewness classifies the second image as more heterogeneous than image 1 in 45.2% of the simulations, kurtosis in 68.9%, entropy in 44.2%, and our proposed NBTH method in 78% of the simulations. This simulation study shows that our proposed method performs better than the competitor methods in terms of ranking images by their heterogeneity. The R code for reproducing all results in our simulation studies is available at https://github.com/anieloyan1/NBTH.

4 Prediction of Cancer Survival

In this section we present results on the comparison of tools for prediction of cancer survival using CT scans from 422 patients with NSCLC publicly available on TCIA. We compare the performance of the proposed tumor heterogeneity estimation approach to five existing methods: two commonly used histogram based measures (skewness and kurtosis); a method for estimating texture features using grey level co-occurrence and grey level run-length features (Haralick et al., 1973); and the method described by Aerts et al. (2014) that combines many existing estimation approaches. Let T_i denote the random variable corresponding to time to the event of death for the ith participant, i.e. the survival time, and C_i denote the time to censoring, i.e. participant leaving the study, or end of study, for $i = 1, \ldots, N$. Then the death status $\delta_i = I(T_i \le C_i)$. For participant i, we observe the outcome pair $\tilde{T}_i = min(T_i, C_i)$ and δ_i. In the dataset, the average observed survival time $\tilde{T}_i$ (including the participants where death was not observed) was 538.6 days (with sd = 417.1, and range = [10, 2165]), the censoring rate was 42%. We are interested in modeling the hazard function $h_i(t)$ defined as the instantaneous risk of occurrence of death, based on, among other features, the novel measure of tumor heterogeneity.

$$h_i(t) = \lim_{\Delta t \to 0}\left(\frac{P(t \le T_i \le t + \Delta t \mid T_i \ge t)}{\Delta t}\right) \tag{8}$$

To compare our tumor heterogeneity estimation approach to the method proposed by Aerts et al. (2014), we generated 431 radiomic features per patient following Aerts et al. (2014). We describe these procedures here briefly and refer the reader to Aerts et al. (2014) for more details. These radiomic features comprise 14 histogram features related to the first order statistics defined as $M_{i1}, \ldots, M_{im}$, for $m = 14$ in Section 2.3. In addition, 8 features are computed related to tumor shape and size defined as $S_{i1}, \ldots, S_{is}$, for $s = 8$. Thirty three features are related to tumor texture computed to quantify intratumor texture differences that can be observed on the CT image and defined as $T_{i1}, \ldots, T_{it}$, for $t = 33$. The remaining 376 are based on wavelet decompositions of the CT image. We applied the Daubechies orthonormal compactly supported wavelet of length 8 least asymmetric family on the original CT images, resulting in 8 decompositions (Daubechies, 1988). For each of the decompositions, the histogram and texture features were computed (47 features in total for each decomposition), leading to a total of 376 wavelet features denoted as $W_{i1}, \ldots, W_{iw}$, where $w = 376$. When computing radiomic features that require discretization of the voxel intensities within the CT image, we used 64 intensity levels. The corresponding vector of tumor heterogeneity features computed following the procedures proposed by Aerts et al. (2014) is denoted as $\widehat{H}_{i,a} = (M_{i1}, \ldots, M_{im}, S_{i1}, \ldots, S_{is}, T_{i1}, \ldots, T_{it}, W_{i1}, \ldots, W_{iw})$, where $\widehat{H}_{i,a} \in R^{431}$.

Let $X_i^1, \ldots, X_i^6$ define the clinical variables age, sex, overall cancer stage, clinical N stage, clinical T stage, and histology for patient i. We consider three sets of variables for prediction of cancer survival as follows: Set 1 - clinical variables $X_i^1, \ldots, X_i^6$; Set 2 - clinical variables $X_i^1, \ldots, X_i^6$, and the subject specific intensity histogram skewness defined by M_{i3}; Set 3 - clinical variables $X_i^1, \ldots, X_i^6$, and the subject specific intensity histogram kurtosis defined by M_{i4}; Set 4 - clinical variables $X_i^1, \ldots, X_i^6$, and the sequence of grey level feature vector that we define by $\widehat{H}_{i,gl}$; Set 5 - clinical variables $X_i^1, \ldots, X_i^6$, and the competitor set of radiomic features $\widehat{H}_{i,a}$ used to estimate tumor heterogeneity in an earlier study Aerts et al. (2014); Set 6 - clinical variables $X_i^1, \ldots, X_i^6$ and our proposed NBTH estimate of tumor heterogeneity $\widehat{H}_{i,NBTH}$. As our first model for association of tumor heterogeneity and hazard function we first consider the Cox Proportional Hazards model.

$$\text{Set 1: } h_i(t) = h_0(t)e^{X_i^T\beta}$$

$$\text{Set2:} h_i(t) = h_0(t)e^{X_i^T\beta + M_{i3}\beta_2}$$

$$\text{Set3:} h_i(t) = h_0(t)e^{X_i^T\beta + M_{i4}\beta_3}$$

$$\text{Set4:} h_i(t) = h_0(t)e^{X_i^T\beta + \widehat{H}_{i,gl}^T\beta_4}$$

$$\text{Set5:} h_i(t) = h_0(t)e^{X_i^T\beta + \widehat{H}_{i,a}^T\beta_5}$$

$$\text{Set6:} h_i(t) = h_0(t)e^{X_i^T\beta + \widehat{H}_{i,NBTH}^T\beta_6}$$

where $h_i(t)$ is the hazard function for patient i defined in (8) and $h_0(t)$ is the baseline hazard. We tested whether the proportional hazard assumption of a Cox Proportional Hazards model was justified in all models using a weighted residual-based method proposed by (Grambsch and Therneau, 1994) and found that the assumption was justified based on the global test in all models.

A direct application of the Cox Proportional Hazards model can be problematic when using variables in Set 5 (i.e. the model using the features estimated according to the procedure by Aerts et al. (2014) along with clinical variables) as the number of radiomic features is large compared with the number of participants. Additionally, according to the analysis of the data performed in this study we found that some of the features are highly correlated. Hence, to reduce the dimension of the radiomic features we apply principal component analysis (PCA). We compute the PCA decomposition of the radiomic feature matrix and define the first p principal components describing 95% of variability in the data by $V_{i1}, \ldots, V_{ip}$. These principal components are used in addition to the clinical variables as inputs to the Cox Proportional Hazards model whenever we refer to using the features in Set 5 described above. This approach is commonly used in high dimensional data analysis and is referred to as Principal Components Regression Jolliffe (1982). This step is not necessary for the other competitor models as well as our proposed modeling approach (using features from Set 6), as the number of features is much smaller.

In addition to the Cox Proportional Hazards model, we consider several other approaches for modeling the association of tumor heterogeneity and the hazard function. In particular, we implement regularized regression approaches to relate the clinical and radiomic feature data with survival (using features from Sets 1–6) using the glmnet package in the R software. Ridge, Lasso, and Elastic-Net penalization within the Generalized linear models (GLM, Friedman et al. (2010)) are used for predictor Sets 1–6. Tuning parameter selection is performed using 10-fold cross-validation. The negative logarithm of the partial likelihood is penalized using the following penalization function.

$$\lambda\left[(1-\alpha)\|\beta\|_2^2/2+\alpha\|\beta\|_1\right]$$

where $\|\cdot\|_2$ describes the L_2 norm function and $\|\cdot\|_1$ is the L_1 norm. When $\alpha = 0$ we obtain the ridge regression penalization, $\alpha = 1$ gives the Lasso penalty, while other values of α result in Elastic-Net. All predictive models are trained on a subset of the data and predictions are computed on a hold out (test) set.

In addition to the Cox Proportional Hazards and regularized regression modeling, we implement machine learning algorithms for prediction of survival using each set of variables. Random survival forest (RSF), proposed by Ishwaran et al. (2008), was used to estimate cancer survival for each set of variables. RSF is an ensemble tree method that extends Breiman (2001) random forests for modeling right-censored survival data. In the first stage of the algorithm, a randomly drawn bootstrap sample of the training data is used to grow a binary tree. At the second stage, a randomly selected subset of the model variables is chosen as candidate variables for splitting each node of the tree. The tree is grown using a combination of averaging over trees and randomization. The next method we implement is the support vector machines (SVM) for survival data (Van Belle et al. (2007), Van Belle et al. (2011)). In this approach, independently right censored and observed failure time data are modeled using a health index which is a proxy between the covariates observed for the subject and their observed failure times. The risk is defined using the concordance index measuring the discriminative power of the health index. Third, we use a k-nearest neighbors survival probability prediction method (BNN) (Lowsky et al., 2013), where the survival curve prediction is constructed using a weighted Kaplan-Meier estimator based on the K most similar training observations. Finally, we apply boosting for high-dimensional time-to-event data for prediction of cancer survival (Binder et al. (2009), Binder et al. (2013)). In this approach, a boosting approach is implemented for fitting a proportional subdistribution hazards model, i.e. the instantaneous risk of having an event in the absence of competing events. This model can incorporate a large number of radiomic features (coded as optional variables), while also taking clinical covariates (coded as mandatory variables) into account.

Figure 6 presents time-dependent ROC (Heagerty et al., 2000) curves describing the results of predictive accuracy of survival based on the pairings of predictors used in the study as defined in this section. Each plot in Figure 6 corresponds to the predictive algorithm: a. Cox Proportional Hazards model with PCA, b. GLM with Lasso penalty, c. GLM with Ridge penalty, d. GLM with Elastic-Net penalty, e. Random Survival Forest, f. survival SVM, g. Boosting, h. BNN. The time-dependent ROC curves were computed for 3-year survival. The colors of the ROC curves correspond to the sets of variables used in each prediction method. Table 2 presents the area under the curve (AUC) values corresponding to each 3-year survival ROC curve in Figure 6. The implementation of survival SVM when using clinical variables and grey level features did not converge, hence the “NA” entry in Table 2. In this example, the method with the best performance in terms of maximizing the AUC is our proposed NBTH estimation procedure used in a GLM model with the Elastic-Net penalty.

Finally, note, that the above assessment of predictive accuracy was based on a holdout method, where model accuracy was assessed for only a single testing set. Thus, to obtain a well-rounded understanding we in addition perform a more rigorous assessment of predictive accuracy using 10-fold cross-validation. Specifically, we divide the dataset into 10 subsets (folds). In each of the iterations of the cross-validation analysis we train the algorithm on 9 of the folds and apply the resulting predictive algorithm to obtain estimates of survival for the 10th fold, where each of the folds is used once as a test set. For each iteration we compute the AUCs based on time-dependent ROC curves. The average AUCs, along with their standard deviations, are presented in Table 3 and the boxplots comparing each model using cross-validation AUCs are shown in Figure 7. The results in Table 3 and Figure 7 indicate, that the incorporation of proposed NBTH features along with clinical information in a GLM model with the ridge penalty results in better predictive accuracy on average, as measured by AUC in this cross-validation analysis. In most cases, using clinical information alone results in a poor predictive performance compared with approaches incorporating texture features such as those in Set 4, 5, and 6. Similarly, when using skewness or kurtosis along with clinical variables the models perform poorly compared to the methods including texture based features. The models incorporating texture features estimated by grey level co-occurrence and grey level run-length variables in addition to clinical information perform better than using the clinical information alone, or clinical variables combined with skewness or kurtosis.

5 Summary and Conclusions

As part of radiological assessment of cancer patients, images of different modalities such as CT, PET, or MRI are often collected and stored for use in disease diagnostics. The calculation of tumor heterogeneity has been based on these images and has been used for a wide spectrum of medical purposes including, but not limited to development of targeted therapies, studying disease progression, response to treatments, and in general the prediction of clinical outcomes such as survival. In this work, we investigated the estimation of tumor heterogeneity, based on medical imaging of patients with lung cancer. We discussed the existing approaches to estimation of tumor heterogeneity in radiological imaging of cancer, including histogram based approaches, and proposed a novel approach based on modeling the intensity profiles using MRFs to take into account the spatial structure of the image. Our proposed method is based on modeling the voxel intensities using a Gaussian mixture density with priors on the mixture weights to take into account the intensities of neighboring voxels. The mixture means and standard deviations are fixed, while the mixture weights are estimated using the EM-algorithm with Langrangian multipliers to model the constraints on the parameters. We compared the resulting parameter estimation to the k-means approach and GMM in simulation studies. We used the resulting MRF labeling to compute summary features describing tumor heterogeneity. Having defined the novel estimator of tumor heterogeneity, we modeled cancer survival as a function of the newly defined estimator. Further, we compared the predictive accuracy of the existing definitions to our proposed estimator when predicting cancer survival. The performance of the methods was evaluated using time-dependent ROC analysis for a cohort of lung cancer patients, followed by cross-validation. Our newly proposed estimator outperformed existing approaches in terms of predictive accuracy.

In this work, we assumed that each patient has one tumor for simplicity of notation and since the participants in the motivating NSCLC data used in this article only had one tumor. If two or more cancer tumors are present in the CT scan, then the proposed methods can be directly applied to estimate tumor-specific heterogeneity measures for each tumor of each study participant. Further investigation is necessary to find how the association of the resulting tumor specific heterogeneity measures and survival should be modeled. Another limitation of the current study is its focus on a single kind of cancer (lung). Further research needs to be conducted based on data from patients with other types of cancer (e.g. glioblastoma), to investigate the applicability of the proposed tumor heterogeneity estimation approach. In this work, our goal was to estimate a small number of image features that described tumor heterogeneity. When predicting cancer survival, future work may consider using deep learning methods to estimate survival using the full 3-dimensional image. In addition, we evaluated the stability of the proposed MRF-GMM approach for voxel labeling and the tumor heterogeneity estimator in the presence of noise in simulation studies. Future work may examine the stability of the proposed survival prediction algorithms in other similar cancer imaging data, as well as generalizability of the proposed estimator for other imaging modalities (e.g. MRI, PET) and other types of cancer.

An important advantage of our proposed definition is that it extends histogram based estimation techniques for tumor heterogeneity to include the spatial structure of the tumor. The variables entering the proposed estimator are based on an MRF labeling approach to find areas of the tumor with similar intensities. During clustering, rather than artificially fixing the number of clusters we choose the number of labels using a novel hypothesis testing-based approach and include the resulting cluster-based feature estimates along with clinical variables in machine learning for building the predictive model.

An intriguing area of future research is the use of the proposed estimator in targeted therapy development. For example, if location-specific genomic information is available from a tumor, we can identify associations of the estimated local moments and cluster features with genomic information and use these associations for targeted therapy. In particular, these associations can be used to develop predictions on what types of targeted therapies could be successfully used for treatment of a new patient based solely on their imaging. Although not considered in this article, this example shows the potential of the proposed methods in targeted therapy development.

Acknowledgments

The project described is supported in part by the NIH grant COBRE P20GM103645 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Aerts H, Rios Velazquez E, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, and Lambin P. Data from nsclc-radiomics. Cancer Imaging Archive, 2015.

Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications, 5, 2014.

Andersen PK and Gill RD. Cox’s regression model for counting processes: a large sample study. The annals of statistics, pages 1100–1120, 1982.

Binder H, Allignol A, Schumacher M, and Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics, 25(7):890–896, 2009. [PubMed: 19244389]

Binder H, Benner A, Bullinger L, and Schumacher M. Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures. Statistics in medicine, 32(10):1778–1791, 2013. [PubMed: 22786659]

Blekas K, Likas A, Galatsanos NP, and Lagaris IE. A spatially constrained mixture model for image segmentation. IEEE transactions on Neural Networks, 16(2):494–498, 2005. [PubMed: 15787156]

Breiman L. Random forests. Machine learning, 45(1):5–32, 2001.

Brooks FJ and Grigsby PW. Quantification of heterogeneity observed in medical images. BMC medical imaging, 13(1):7, 2013. [PubMed: 23453000]

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging, 26(6):1045–1057, 2013. [PubMed: 23884657]

Daubechies I. Orthonormal bases of compactly supported wavelets. Communications on pure and applied mathematics, 41(7):909–996, 1988.

Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, Ganeshan B, Miles KA, Cook GJ, and Goh V. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights into imaging, 3(6):573–589, 2012. [PubMed: 23093486]

Dempster AP, Laird NM, and Rubin DB. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

Dhanachandra N, Manglem K, and Chanu YJ. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54:764–771, 2015.

Eloyan A and Ghosh SK. Smooth density estimation with moment constraints using mixture distributions. Journal of nonparametric statistics, 23(2):513–531, 2011.

Friedman J, Hastie T, and Tibshirani R. The elements of statistical learning. Springer series in statistics New York, 2001.

Friedman J, Hastie T, and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010. [PubMed: 20808728]

Grambsch PM and Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3):515–526, 1994.

Gutierrez DR, Awwad A, Meijer L, Manita M, Jaspan T, Dineen RA, Grundy RG, and Auer DP. Metrics and textural features of mri diffusion to improve classification of pediatric posterior fossa tumors. American Journal of Neuroradiology, 35(5):1009–1015, 2014. [PubMed: 24309122]

Haralick RM, Shanmugam K, et al. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, pages 610–621, 1973.

Heagerty PJ, Lumley T, and Pepe MS. Time-dependent roc curves for censored survival data and a diagnostic marker. Biometrics, 56(2):337–344, 2000. [PubMed: 10877287]

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS, et al. Random survival forests. The annals of applied statistics, 2(3):841–860, 2008.

Jolliffe IT. A note on the use of principal components in regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 31(3):300–303, 1982.

Just N. Improving tumour heterogeneity mri assessment with histograms. British journal of cancer, 111(12):2205–2213, 2014. [PubMed: 25268373]

Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282):457–481, 1958.

Li SZ. Markov random field modeling in image analysis. Springer Science & Business Media, 2009.

Lowsky DJ, Ding Y, Lee DK, McCulloch CE, Ross LF, Thistlethwaite JR, and Zenios SA. Ak-nearest neighbors survival probability prediction method. Statistics in medicine, 32(12):2062–2069, 2013. [PubMed: 23653217]

Meng K and Eloyan A. Principal manifolds: A framework using sobolev spaces and model complexity selection using mixture densities. arXiv preprint arXiv:1711.06746, 2017.

Nevo D, Zucker DM, Tamimi RM, and Wang M. Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data. Statistics in medicine, 35(30):5686–5700, 2016. [PubMed: 27558651]

Nguyen TM and Wu QJ. Fast and robust spatially constrained gaussian mixture model for image segmentation. IEEE transactions on circuits and systems for video technology, 23(4):621–635, 2012.

Ni Y, Stingo FC, Ha MJ, Akbani R, and Baladandayuthapani V. Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. Journal of the American Statistical Association, 114(525):48–60, 2019. [PubMed: 31178611]

O'Sullivan F, Roy S, O'Sullivan J, Vernon C, and Eary J. Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with fdg-pet. Biostatistics, 6(2):293–301, 2005. [PubMed: 15772107]

Peng S-L, Chen C-F, Liu H-L, Lui C-C, Huang Y-J, Lee T-H, Chang C-C, and Wang F-N. Analysis of parametric histogram from dynamic contrast-enhanced mri: application in evaluating brain tumor response to radiotherapy. NMR in Biomedicine, 26(4):443–450, 2013. [PubMed: 23073840]

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

Rao CR. Diversity and dissimilarity coefficients: a unified approach. Theoretical population biology, 21(1):24–43, 1982.

Roberts T, Newell M, Auffermann W, and Vidakovic B. Wavelet-based scaling indices for breast cancer diagnostics. Statistics in medicine, 36(12):1989–2000, 2017. [PubMed: 28226399]

Van Belle V, Pelckmans K, Suykens J, and Van Huffel S. Support vector machines for survival analysis. In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), pages 1–8, 2007.

Van Belle V, Pelckmans K, Van Huffel S, and Suykens JA. Improved performance on high-dimensional survival data by application of survival-svm. Bioinformatics, 27(1): 87–94, 2011. [PubMed: 21062763]

Wagner S and Wagner D. Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

Wang Q. Gmm-based hidden markov random field for color image and 3d volume segmentation. arXiv preprint arXiv:1212.4527, 2012.

Xu Y, Müller P, Yuan Y, Gulukota K, and Ji Y. Mad bayes for tumor heterogeneity—feature allocation with exponential family sampling. Journal of the American Statistical Association, 110(510):503–514, 2015. [PubMed: 26170513]

Zhang L and Ji Q. Image segmentation with a unified graphical model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8):1406–1425, 2009.

Zhang Y, Brady M, and Smith S. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging, 20(1):45–57, 2001. [PubMed: 11293691]

Zhu X, Ghahramani Z, and Lafferty JD. Semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the 20th International conference on Machine learning (ICML-03), pages 912–919, 2003.

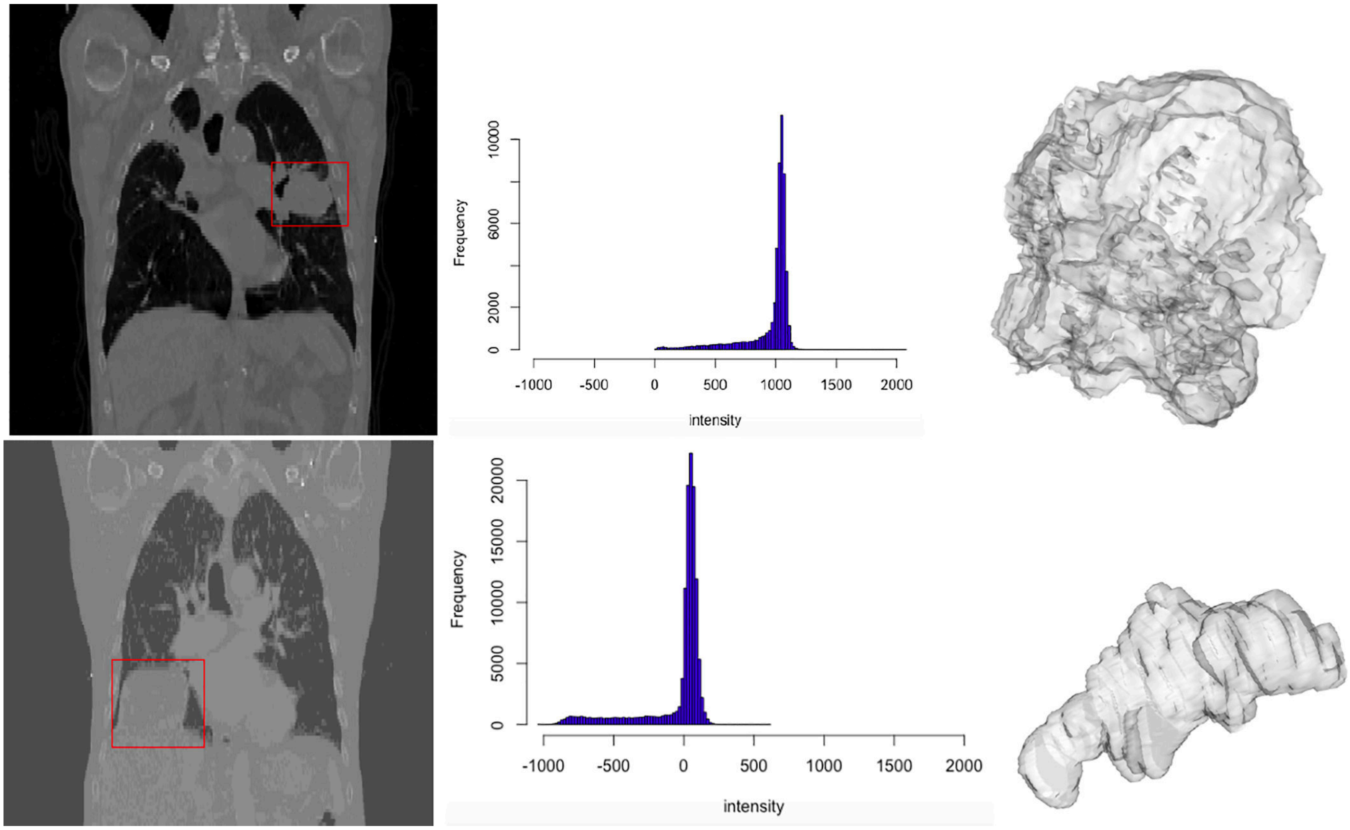

Figure 1:

Left column: 2D slices of CT scans of two patients. The tumor is enclosed in a red box. Middle column: histograms showing the intensities of tumor voxels. Right column: 3D surface renditions of respective tumors.

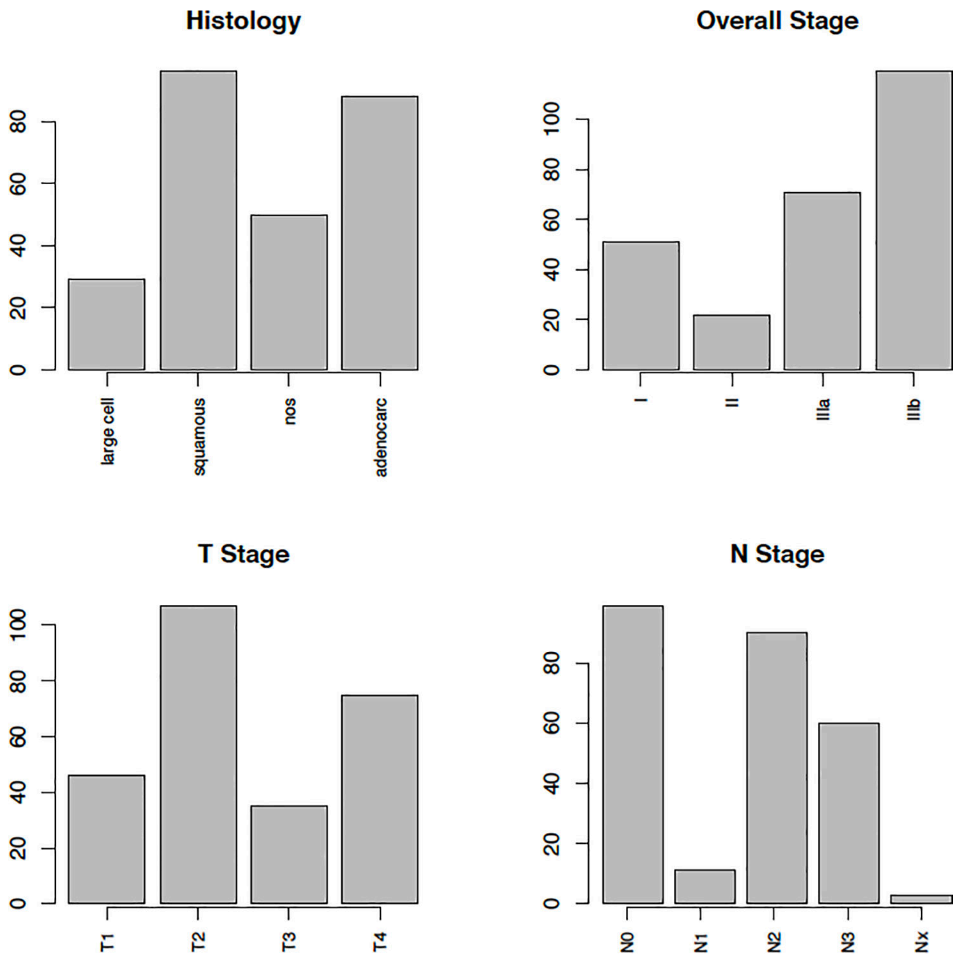

Figure 2:

Barplots showing the distribution of clinical variables. Histology provides a description of the tumor based on the abnormality of tumor cells observed using a microscope. Overall stage of cancer indicated the size of the tumor and its spread by a number I, II, IIIa, and IIIb, where a higher number implies larger and more spread cancer. T1, T2, T3, and T4 refer to the size and extent of the tumor. The higher the number the larger or more extensive the tumor. N1, N2, and N3 refer to the number and location of lymph nodes that contain cancer (the higher the number the more lymph nodes), N0 indicates there are no nearby lymph nodes, NX implies that cancer in nearby lymph nodes is not measurable.

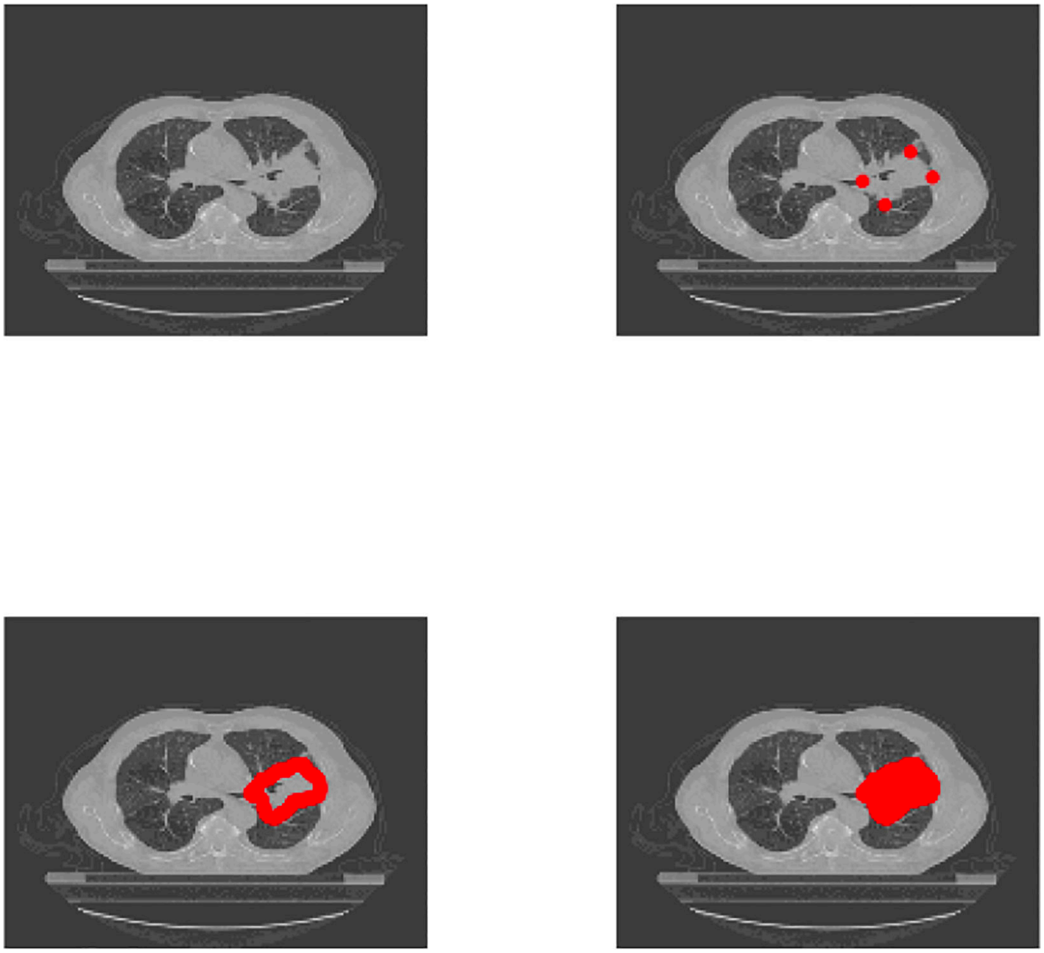

Figure 3:

Top left: An axial slice of the CT for one patient. Top right: Four of the tumor surface vertices (in red) marked by the radiologist. Bottom left: All tumor surface points marked by the radiologist. Bottom right: The full area of the tumor identified by interpolation of surface points and identification of tumor interior points.

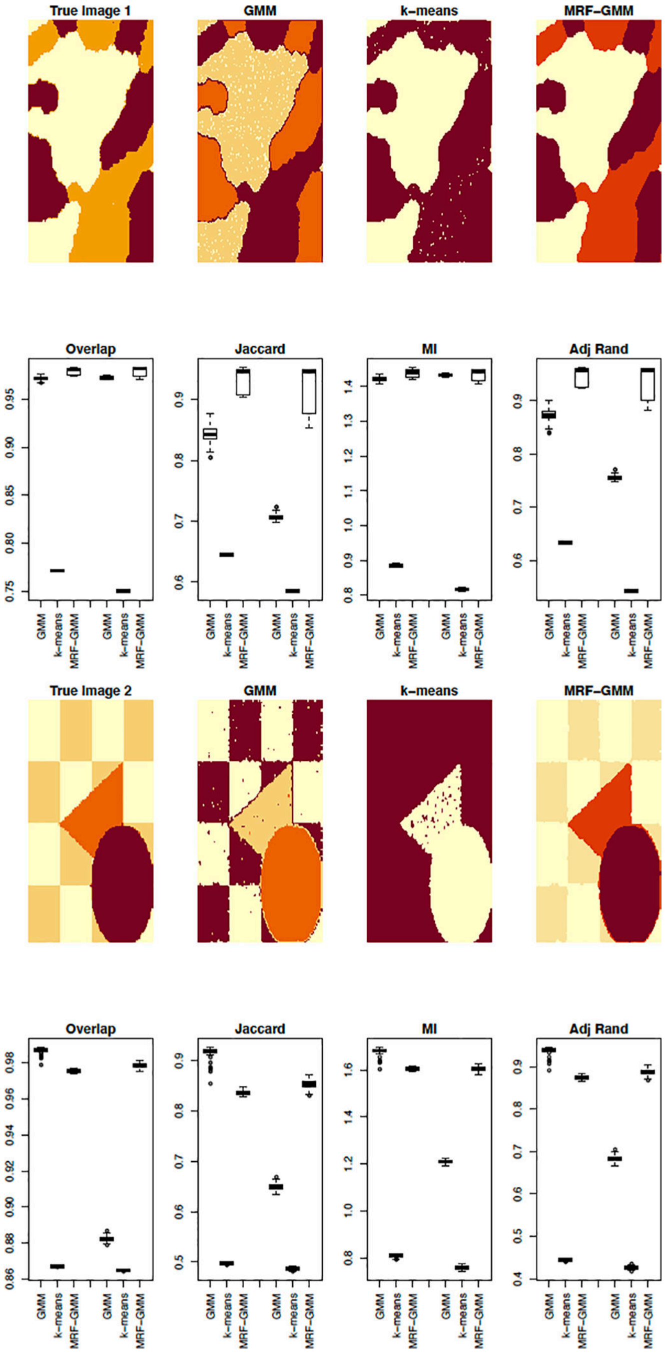

Figure 4:

Results of clustering the voxel intensities using k-means, GMM, and our proposed MRF-GMM. The true images as well as the estimated cluster segmentations by each of the three algorithms are shown on rows 1 and 3. Each of these rows are followed by a row of boxplots showing the results of the four cluster comparison measures (overlap, Jaccard index, mutual information (MI) and adjusted Rand index) for 100 runs of the simulations. Within each boxplot figure, the first three boxplots correspond to the first added noise setting ($\sigma = 1$ and $\sigma = 2$ for true images 1 and 2 correspondingly) while the second set of three boxplots corresponds to the second added noise setting ($\sigma = 2$ and $\sigma = 3$ for true images 1 and 2 correspondingly).

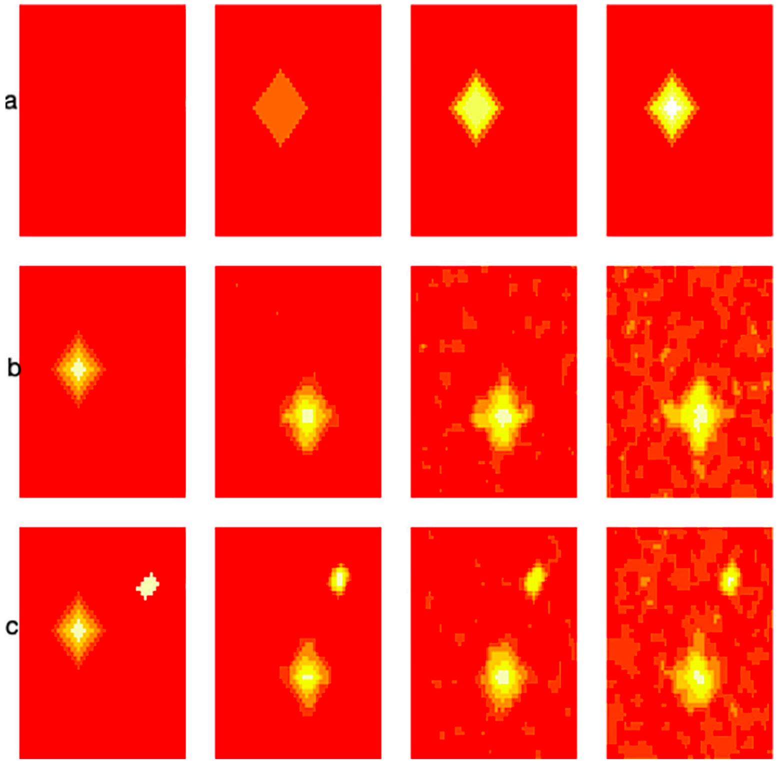

Figure 5:

True images used in the simulations to evaluate ranking of image heterogeneity.

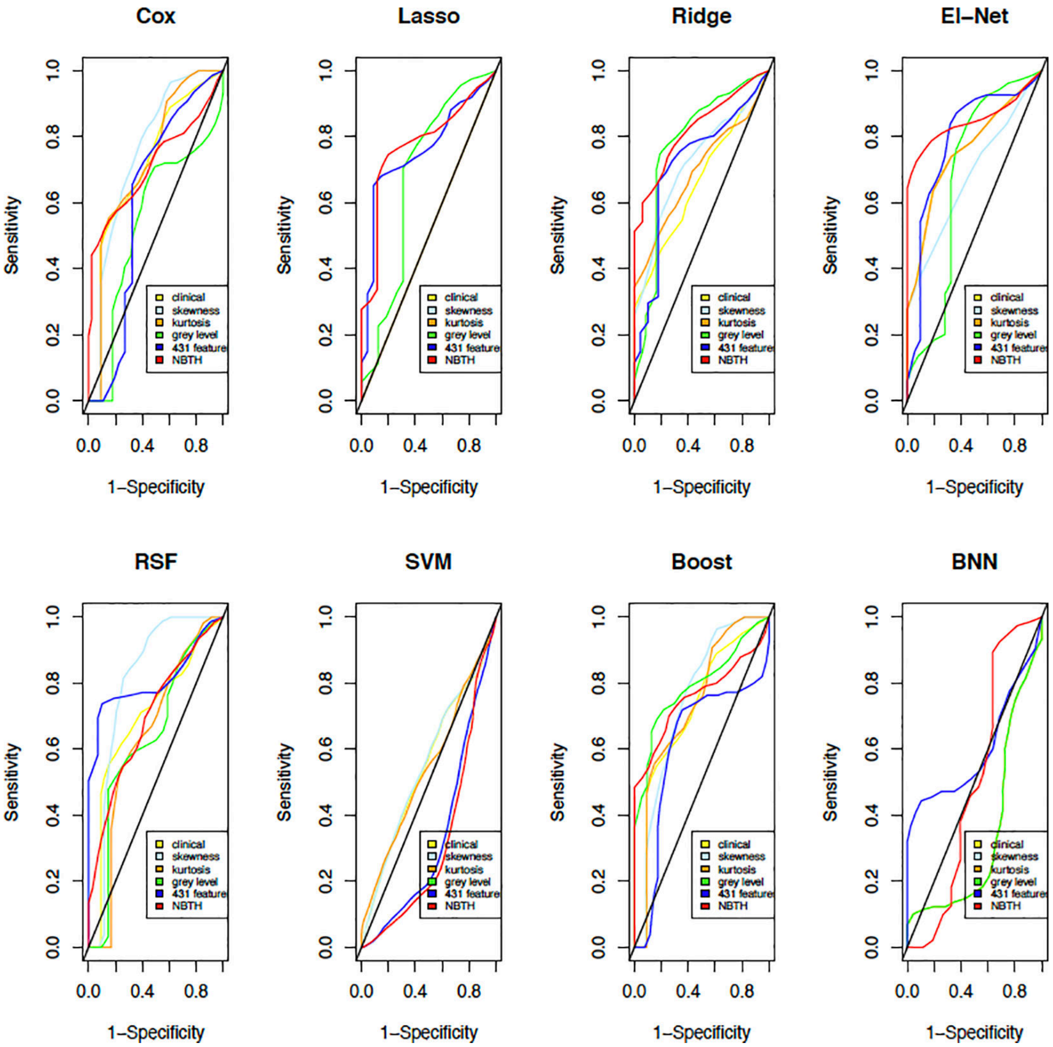

Figure 6:

ROC curves comparing the six predictive methods using the proposed collections of tumor heterogeneity measures for 3-year survival.

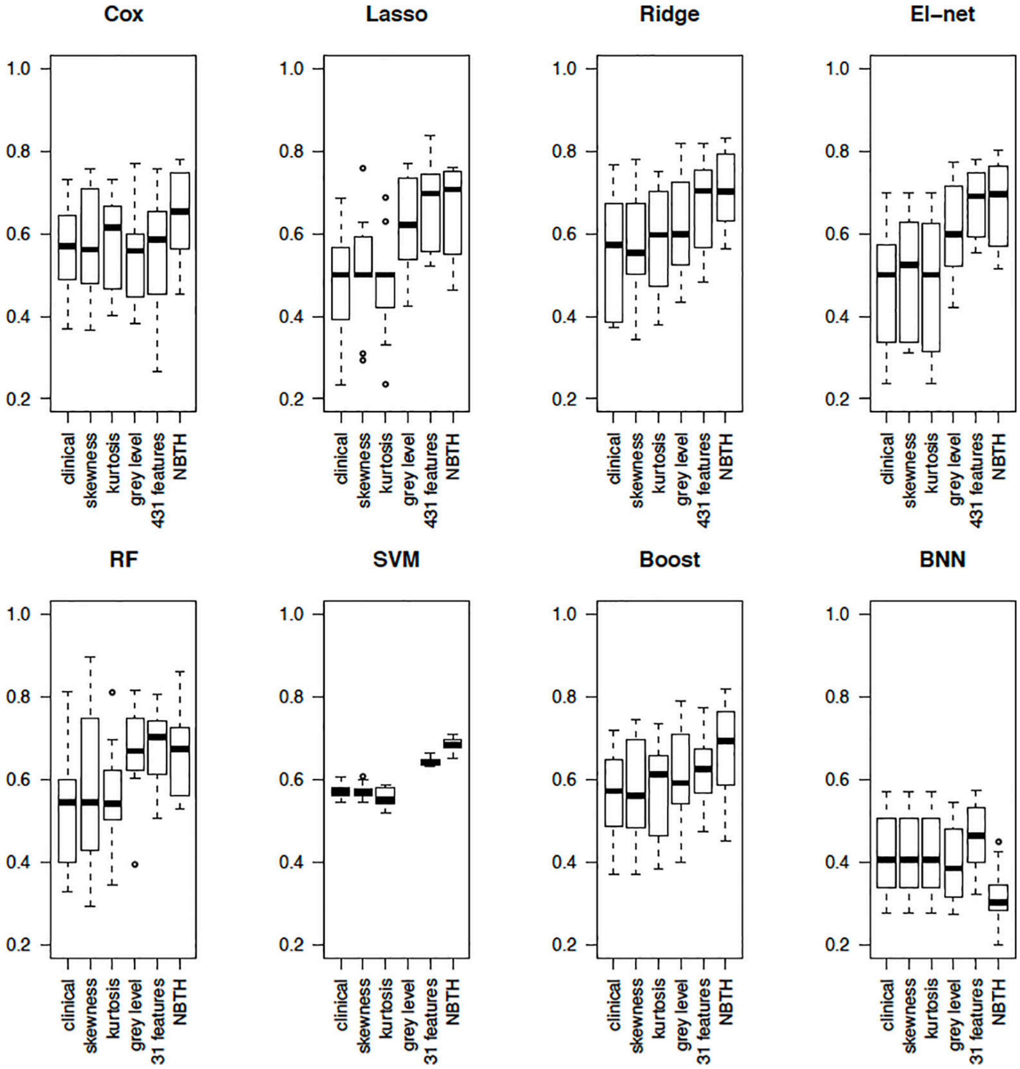

Figure 7:

Boxplots of AUCs from 10-fold cross-validation for each model and the corresponding estimation method for 3-year survival.

Table 1:

Summary of cluster comparison measures given as mean (SD*100) using several measures of cluster comparisons OL - overlap, J - Jaccard Index, MI - Mutual Information, ARI - Adjusted Rand Index.

	$\sigma = 1$, image 1			$\sigma = 2$, image 1		
	GMM	k-means	MRF-GMM	GMM	k-means	MRF-GMM
OL	0.97	0.77	**0.98**	0.97	0.75	**0.98**
	(0.2)	(0.04)	(0.38)	(0.09)	(0.03)	(0.43)
J	0.84	0.64	**0.93**	0.71	0.59	**0.92**
	(1.39)	(0.11)	(2.08)	(0.54)	(0.06)	(3.6)
MI	1.42	0.88	**1.44**	1.43	0.82	**1.44**
	(0.66)	(0.31)	(1.19)	(0.29)	(0.29)	(1.38)
ARI	0.87	0.63	**0.95**	0.76	0.54	**0.94**
	(1.18)	(0.13)	(1.68)	(0.49)	(0.07)	(2.94)
	$\sigma = 2$, image 2			$\sigma = 3$, image 2		
	GMM	k-means	MRF-GMM	GMM	k-means	MRF-GMM
OL	**0.99**	0.87	0.98	0.88	0.86	**0.98**
	(0.13)	(0.02)	(0.07)	(0.14)	(0.03)	(0.13)
J	**0.92**	0.50	0.84	0.65	0.21	**0.80**
	(1.03)	(0.13)	(0.43)	(0.65)	(0.21)	(0.8)
MI	**1.68**	0.81	1.60	1.21	0.76	**1.61**
	(1.36)	(0.53)	(0.6)	(0.74)	(0.74)	(1.0)
ARI	**0.94**	0.45	0.87	0.68	0.43	**0.89**
	(0.08)	(0.21)	(0.36)	(0.73)	(0.36)	(0.66)

Table 2:

Area under the curve for each variable set and the corresponding estimation method for 3-year survival.

	Cox	**Lasso**	**Ridge**	**El-Net**	**RSF**	**SVM**	**Boost**	**BNN**
Set 1	0.72	0.50	0.66	0.76	0.70	0.55	0.72	0.34
Set 2	0.76	0.50	0.72	0.68	0.80	0.55	0.75	0.34
Set 3	0.73	0.50	0.69	0.76	0.64	0.54	0.73	0.34
Set 4	0.54	0.68	0.78	0.67	0.63	NA	0.79	0.34
Set 5	0.61	0.76	0.72	0.79	0.81	0.34	0.61	0.60
Set 6	0.71	0.78	0.82	**0.85**	0.70	0.32	0.76	0.50

Table 3:

Average (standard deviation) AUCs from 10-fold cross-validation for each variable set and the corresponding estimation method for 3-year survival.

	Cox	Lasso	Ridge	El-Net	RSF	SVM	Boost	BNN
Set 1	0.57	0.49	0.57	0.48	0.54	0.57	0.57	0.42
	(0.12)	(0.13)	(0.15)	(0.14)	(0.15)	(0.02)	(0.12)	(0.09)
Set 2	0.57	0.51	0.57	0.50	0.58	0.57	0.57	0.42
	(0.14)	(0.14)	(0.15)	(0.15)	(0.19)	(0.02)	(0.13)	(0.09)
Set 3	0.57	0.48	0.59	0.46	0.56	0.56	0.57	0.42
	(0.11)	(0.13)	(0.13)	(0.16)	(0.13)	(0.02)	(0.12)	(0.09)
Set 4	0.55	0.63	0.62	0.61	0.66	NA	0.60	0.40
	(0.12)	(0.12)	(0.13)	(0.12)	(0.12)	NA	(0.12)	(0.09)
Set 5	0.56	0.67	0.67	0.67	0.67	0.36	0.62	0.46
	(0.15)	(0.11)	(0.11)	(0.08)	(0.09)	(0.01)	(0.09)	(0.08)
Set 6	0.65	0.66	**0.71**	0.68	0.66	0.32	0.68	0.32
	(0.11)	(0.11)	(0.09)	(0.10)	(0.11)	(0.02)	(0.12)	(0.07)