Assessing the scale of tumor heterogeneity by complete hierarchical segmentation of MRI

Michael F Gensheimer¹, Douglas S Hawkins², Ralph P Ermoian¹ and Andrew D Trister³

¹ Department of Radiation Oncology, University of Washington Medical Center, 1959 NE Pacific St. Box 356043, Seattle, WA 98115, USA
² Division of Hematology/Oncology, Seattle Children’s Hospital, MB.8.501 – Hematology–Oncology, 4800 Sand Point Way NE, Seattle, WA 98105, USA
³ Sage Bionetworks, 1100 Fairview Avenue N, Mailstop M1-C108, Seattle, WA 98109, USA

E-mail: trister@alumni.upenn.edu

Received 23 May 2014, revised 3 November 2014
Accepted for publication 18 November 2014
Published 9 January 2015

Abstract

In many cancers, intratumoral heterogeneity has been found in histology, genetic variation and vascular structure. We developed an algorithm to interrogate different scales of heterogeneity using clinical imaging. We hypothesize that heterogeneity of perfusion at coarse scale may correlate with treatment resistance and propensity for disease recurrence. The algorithm recursively segments the tumor image into increasingly smaller regions. Each dividing line is chosen so as to maximize signal intensity difference between the two regions. This process continues until the tumor has been divided into single voxels, resulting in segments at multiple scales. For each scale, heterogeneity is measured by comparing each segmented region to the adjacent region and calculating the difference in signal intensity histograms. Using digital phantom images, we showed that the algorithm is robust to image artifacts and various tumor shapes. We then measured the primary tumor scales of contrast enhancement heterogeneity in MRI of 18 rhabdomyosarcoma patients. Using Cox proportional hazards regression, we explored the influence of heterogeneity parameters on relapse-free survival. Coarser scale of maximum signal intensity heterogeneity was prognostic of shorter survival (p = 0.05). By contrast, two fractal parameters and three
Haralick texture features were not prognostic. In summary, our algorithm produces a biologically motivated segmentation of tumor regions and reports the amount of heterogeneity at various distance scales. If validated on a larger dataset, this prognostic imaging biomarker could be useful to identify patients at higher risk for recurrence and candidates for alternative treatment.

Keywords: tumor heterogeneity, biomarkers, image segmentation, magnetic resonance imaging

(Some figures may appear in colour only in the online journal)

1. Introduction

For many cancer types, evidence has emerged that biological heterogeneity exists within tumors and between different tumor deposits in the same patient (Gatenby et al 2011). This multifaceted heterogeneity includes genetic variation of malignant cells and differences in vascular structure in various tumor regions. Contrast-enhanced T1-weighted magnetic resonance imaging (MRI) and the related tool of dynamic contrast enhanced (DCE) MRI may be useful ways to interrogate heterogeneity within tumors. These images capture features such as blood flow, vascular permeability, and tissue necrosis. Cancers with a more heterogeneous appearance on contrast-enhanced T1-weighted MRI may have a poorer prognosis (Ahmed et al 2013).

An additional facet of intratumoral heterogeneity is the physical scale of that heterogeneity. Tumors with coarse- or large-scale heterogeneity may behave differently from those with fine- or small-scale heterogeneity. There are emerging molecular data to support this hypothesis. A recent study in glioblastoma showed that in some patients’ tumors, different regions of the tumor used different cell surface receptors to drive growth; in other patients’ specimens the receptors were distributed more evenly throughout the tumor (Snuderl et al 2011). Another feature that can be present on different scales is necrosis, or fluid-filled dying areas of the tumor. Areas of necrosis can range in size from the micron level to macroscopic scales (Brown and Fletcher 2000). A classic study in gliomas showed that signal intensity (SI) on contrast-enhanced imaging correlated with biopsy findings of viable tumor versus necrosis (Kelly et al 1987). The presence of necrosis demonstrates a higher growth rate and propensity for metastasis, and is included in the grading system for a variety of cancers, including soft tissue sarcomas (Brown and Fletcher 2000). We hypothesize that heterogeneity in contrast enhancement on MRI at coarse distance scales correlates with treatment resistance and propensity for disease recurrence.

Other researchers have examined heterogeneity in MR or CT images of tumors and attempted to correlate the amount or scale of heterogeneity with disease characteristics or outcomes. Among the most commonly applied methods are fractal analysis and texture analysis using gray-level co-occurrence matrices (Gatenby et al 2011, Yang and Knopp 2011). These methods rely on assumptions that may not be accurate: fractal analysis assumes the tumor to be self-similar, while texture analysis methods often assume a uniform texture over the whole tumor.

We have developed an alternative approach of heterogeneity analysis that builds from the concept that tumors are composed of biologically distinct regions on multiple scales: Spatial Heterogeneity Analysis by Recursive Partitioning (SHARP). The method relies on two assumptions: (1) each tumor is composed of regions with distinct molecular features and vascular supply, and (2) the biological similarity of two tumor regions can be measured by comparing their SI histograms—the more similar the histograms, the more similar the biological characteristics. The tumor is recursively subdivided, with the dividing lines chosen so as
to maximize SI difference between the two regions. The process is repeated until the tumor has been divided into single voxels. To report average heterogeneity at each physical scale, each region of that scale is compared to the adjacent region by finding the difference in SI histograms.

In this report, we describe the SHARP algorithm and its potential applications. As a motivation of the clinical applications, we also present the results of applying the SHARP algorithm to initial diagnosis MRI of patients with rhabdomyosarcoma for prediction of recurrence risk.

2. Methods

2.1. SHARP algorithm

The SHARP algorithm takes as input a 2D or 3D image of a tumor, and outputs:

(a) a binary tree that divides the tumor into regions;
(b) for each node of the tree, the difference in SI histograms between the two children of that node, as measured by the Earth mover’s distance (EMD);
(c) a list of distance scales (e.g. [32 mm, 16 mm, 8 mm, 4 mm]), and the mean SI heterogeneity at each scale as measured by the mean EMD as calculated in (2)

In practice, the SHARP algorithm has two steps (visual schematic is shown in figure 1, pseudocode is listed in figure 2). First, the tumor is recursively divided in a binary fashion into regions until each voxel has been isolated. The dividing lines are chosen with the goal of maximizing the difference in SI between each pair of regions, using the Normalized Cuts algorithm (Shi and Malik 2000). This algorithm treats the input region as a graph, with one node representing each voxel. The weight between nodes $i$ and $j$ is:

$$w_{ij} = e^{-\frac{(\text{SI}(i) - \text{SI}(j))^2}{\eta^2} - \frac{\|X(i) - X(j)\|^2}{\sigma_X^2}},$$

(1)

where $\text{SI}(i)$ is the signal intensity at node $i$, $X(i)$ is the spatial location of node $i$, and $\eta$ and $\sigma_X$ are constants defining the sensitivity of the partitioning to SI and distance difference between voxels. The input region is divided into two partitions $A$ and $B$ with the goal of maximizing the weight of the edges within each partition, and minimizing the weight of the edges between partitions. An eigenvalue problem is solved so as to maximize $N_{\text{cut}}$, defined as:

$$N_{\text{cut}}(A, B) = 2 - \sum_{u \in A, t \in B} w(u, t).$$

(2)

To reduce execution time and memory requirements, we use the Nyström approximation to the Normalized Cut, which first solves the grouping problem for a random subset of the input voxels and then extends the two groups to include all the input voxels (Fowlkes et al 2004). Because SI on MRI does not have any absolute meaning akin to Hounsfield units on CT, the SI term in equation (1) is normalized by dividing SI by the interquartile range of the whole tumor’s SI (pseudocode line 1). Similarly, the distance term in equation (1) is normalized to the approximate diameter of the input region: $\sigma_X$ is scaled by the diameter of the sphere with equal volume to the input region (pseudocode line 8). This step is performed because when $\sigma_X$ was kept constant despite varying volume of the input region, we found that for very small input regions the distance term in equation (1) was dominated by the SI difference term. The distance normalization ensures that the same arrangement of signal intensities will produce the same partitions regardless of the scale of the input region.
The Normalized Cut algorithm does not guarantee that at each partition step, each of its two partitions is connected (i.e. a path exists from each voxel to every other voxel without leaving the region). We therefore added an evaluation for each partition step: if there are more than two connected components, the components are reassigned so that there are only two connected partitions. Each connected component other than the largest is reassigned to a component to which it is in contact (by a 4-connected neighborhood for 2D images and a 6-connected neighborhood for 3D images).

The output of the recursive partitioning is a binary tree that hierarchically divides the tumor; at the greatest tree depth each node represents a single voxel. We quantify SI heterogeneity.

**Figure 1.** A schematic of the SHARP algorithm. (a) The tumor is recursively divided into subregions, with dividing lines chosen so as to maximize 1) signal intensity difference between the child regions and 2) spatial coherence of each child region. Only the first two divisions are shown. Signal intensity histograms are compared between each pair of child regions using the EMD. Greater dissimilarity of the histograms from each region results in higher EMD. (b) For each pair of child regions, EMD is plotted against the diameter of parent region (normalized to the diameter of the entire tumor). This tumor exhibits more heterogeneity at the coarse distance scale, so EMD is higher at the coarser scale.
Dissimilarity is measured by the EMD (Rubner et al. 2000) between their normalized SI histograms. If the frequency bars on a histogram plot are thought of as piles of dirt, the EMD is the quantity of dirt that must be moved, multiplied by the horizontal distance it is moved, to make the two histograms match. Mathematically it is defined as:

$$\text{EMD} = \sum_{i=1}^{n} |CDF_a(i) - CDF_b(i)|,$$

where histograms $a$ and $b$ each have $n$ bins and $CDF_a$ represents the cumulative distribution function of histogram $a$.

The EMD between identical histograms is 0; EMD increases as the shape of the two histograms diverges, up to a maximum of the number of histogram bins minus one. Compared to simpler metrics such as difference in mean SI, EMD has the advantage of being more sensitive to differences between regions. For example, consider two tumors A and B, each divided into regions #1 and #2. In region A1, half the voxels have SI of 1 and half have SI of 3. In region A2, all voxels have SI of 2. In regions B1 and B2, all voxels have SI of 2. For both tumors, the difference in mean SI between regions is 0. However, the EMD is higher for tumor A than
tumor B: for histogram bin size of 1, $EMD_A = 1$ and $EMD_B = 0$. The EMD corresponds better to the visual appearance of the tumor regions—the two regions of tumor A would appear quite different but the two regions of tumor B would look the same.

For each parent node of the partition tree, we plot EMD against region scale in millimeters (diameter of the sphere with equal volume to the input region) in a scatterplot (figure 1(b)). Data are summarized by binning region scale on a logarithmic scale. From this summary plot, we can extract features such as scale of maximum heterogeneity and slope of heterogeneity versus scale. We can calculate other features from the partition tree, including the area of the tumor with the most/least heterogeneity at each scale, and for a given point in the tumor, the amount of heterogeneity at different scales (figure 3).

SHARP is implemented in MATLAB R2012b (MathWorks, Natick, MA). Source code and an example dataset are available on GitHub [https://github.com/MGensheimer/SHARP]. We use the implementation of Normalized Cuts in Sameer Agarwal’s freely available Spectral Clustering Toolbox (Agarwal 2002). Running time for a tumor with 4000 pixels is 48 s on a Linux system with 16 GB of memory and two Intel Xeon E5620 processors with eight cores.

2.2. Validation studies

We performed validation of the algorithm using three classes of synthetic images: one with more coarse-scale heterogeneity (‘Large blocks’), one with a mix of coarse- and fine-scale heterogeneity (‘Large + small blocks’), and one with more fine-scale heterogeneity (‘Small blocks’) (figure 4(a)). Image size was $64 \times 64$ pixels, with pixel size of $1 \times 1$ mm. Each image was initialized with a uniform value of 0. Then, many rectangles were added, with intensity in the rectangle shifted by an amount randomly sampled from a uniform distribution centered at 0. In the three classes of images, the rectangles of different size are weighted differently. For the ‘Large blocks’ image, the highest weighted rectangles are large, of size $32 \times 64$ mm. For
the ‘Small blocks’ image, the largest rectangles are given a very low weight and the smallest rectangles, of size 1 × 1 mm, are given the highest weight. For the ‘Large + small blocks’ image, all sizes of rectangles have the same weight.

SI was normalized to a scale of 0 to 1, with outlier pixels (intensity below 2nd percentile or above 98th percentile) set to 0 or 1. $C_I$ and $C_X$ were set by running the algorithm on synthetic images and one patient’s MRI, varying $C_I$ and $C_X$, and choosing values that led to partitions that tracked along SI gradients but reliably produced two cohesive regions. The same $C_I$ and $C_X$ values were used throughout this work. 300 samples were used for the

Figure 4. Validation of the SHARP algorithm using synthetic data. (a) Three classes of images are produced, with varying scales of heterogeneity. Representative examples of the three classes are shown here. (b) SHARP algorithm is run on 20 randomly generated images of each class. Points indicate mean value among the 20 images, error bars indicate standard deviation. The ‘Large blocks’ images show greatest heterogeneity at the maximum scale bin of $2^6 = 64$ mm, the full diameter of the image. Conversely, the ‘Small blocks’ images show greatest heterogeneity at the smallest scale bin of 2 mm. When blocks are widely distributed in size (‘Large + small blocks’), heterogeneity at various scales is roughly constant.
Nyström approximation of the Normalized Cut. 16 histogram bins were used when calculating EMD between regions. To report mean EMD at each distance scale, parent region diameter was grouped into six bins ranging from 2 to 64 mm, equally spaced on a logarithmic scale. MATLAB code for generating and analyzing the synthetic images can be found in the `synthetic.m` file in the GitHub repository.

It is important that the algorithm be robust to artifacts such as noise or intensity gradients that can commonly be present in MR images. Therefore we ran simulations where Gaussian white noise or a smoothly varying intensity gradient was added to the three classes of synthetic images. Various strengths of noise or gradient, defined as a percentage of the variation in the initial image, were applied. For example, 10% Gaussian noise meant adding noise with a Gaussian distribution with standard deviation equal to 10% of the standard deviation of the initial image ($\sigma_I$). Since standard deviation is not defined for a linear intensity gradient, for this perturbation the strength of the gradient was measured in percentages of $\sigma_I \cdot 4$ with the justification that 95% of the values of the original normally distributed image lie within a range of $\sigma_I \cdot 4$. So a gradient with 100% strength would create added variation equal to the variation already present in the image.

We measured the effects of the image perturbations on the SHARP algorithm output using intraclass correlation, as recommended by the Quantitative Imaging Biomarkers Alliance (Raunig et al 2015). For each of the three synthetic image classes, and for each level of image perturbation, 20 random images were generated. Then, for each distance scale the EMD was compared between images of the same class with different amounts of noise/gradient (within class variation), and between images of the three different classes (between class variation). Intraclass correlation was calculated using the ICCest function in the ICC R package. An intraclass correlation value of >0.8 indicates good within-group agreement (Mosher et al 2011).

### 2.3. Application to rhabdomyosarcoma tumors

Using the SHARP algorithm, we examined the primary tumor of 18 patients with rhabdomyosarcoma on initial diagnosis MRI, to explore the influence of scale of SI heterogeneity on relapse-free survival (RFS). The Seattle Children’s Hospital Institutional Review Board approved this retrospective study. Eligibility for the study included diagnosis with rhabdomyosarcoma between 2003–2012, pre-treatment contrast-enhanced MRI available in the digital radiology system, and treatment at Seattle Children’s Hospital. All patients were treated with chemotherapy and local treatment (radiation, surgery, or both) according to national standard regimens. Only patients with primary tumor greater than or equal to 2 cm in maximal diameter were included, to ensure that several scales of heterogeneity could be analyzed with SHARP. In total, 18 patients met the eligibility requirements and were included in the analysis. We analyzed the T1-weighted gadolinium-enhanced sequence. In-plane resolution ranged from 0.31 × 0.31 to 1.32 × 1.32 mm. Slice spacing ranged from 3 to 8 mm. Median TR was 568 ms (range 34–850), and median TE was 12 ms (range 3–29).

The tumor border was manually drawn by M.G. on three equally spaced parallel image slices using 3D Slicer version 3 (Pieper et al 2004). Tumor SI was normalized to a scale of 0 to 1. Then, the SHARP algorithm was run in 2D on the tumor portion of each of the three slices. The lists of parent region size and corresponding EMD for the three slices were concatenated. An alternate approach would be to create plots of EMD versus region size for the three slices, followed by an average of the plots. Since tumor size can vary between slices, we instead chose to concatenate the three slices’ lists of region size versus EMD so that if one slice had more regions of a certain scale, each region would influence the final plot equally.
equally spaced on a logarithmic scale. We ran a separate analysis for two bin spacings: powers of 2 (bin centers \{0.25 \text{ mm}, 0.5 \text{ mm}, 1 \text{ mm}, \ldots, 128 \text{ mm}\}), and powers of \(\sqrt{2}\) (bin centers \{0.25 \text{ mm}, 0.353 \text{ mm}, 0.5 \text{ mm}, 0.707 \text{ mm}, \ldots, 128 \text{ mm}\}). To enable unbiased comparison of various sized tumors, distance scale was then divided by the maximal diameter of each tumor.

For calculation of RFS, events were disease relapse (growth of treated lesions or appearance of new lesions) or death. Median follow-up was 35 months. For RFS analysis, patients without an event were censored at last follow-up. Imaging covariates were compared to RFS using univariate Cox proportional hazards regression. We extracted measures expressing the scale of heterogeneity of MRI, rather than the absolute quantity of heterogeneity:

(a) MaxScale: \(\log_2(\text{scale with maximum heterogeneity})\), with heterogeneity defined as mean EMD at that scale

(b) HetSlope: Slope of least squares regression line fitting heterogeneity versus \(\log_2(\text{scale})\)

For both of these measures, scale was normalized to maximal tumor diameter in order to create comparable results between tumors of different diameters. In summary, four heterogeneity measures per patient were produced: scale of maximum heterogeneity and slope of heterogeneity versus scale, each for two scale bin sizes (powers of 2, powers of \(\sqrt{2}\)). These are abbreviated as MaxScale_{2}, HetSlope_{2}, MaxScale_{\sqrt{2}}, and HetSlope_{\sqrt{2}}. The \texttt{coxph} function in R version 3.0.2 was used for survival analysis (R Core Team 2013).

2.4. Comparison with existing methods

We compared the performance of SHARP to three existing methods: spectral analysis using the 2D discrete Fourier transform of the image, fractal analysis and texture analysis using Haralick features.

2.4.1. Spectral analysis. By performing a 2D discrete Fourier transform of an image and then generating a rotationally averaged power spectrum, the amount of signal at different distance scales can be extracted. We implemented a widely used method described in (Balboa and Grzywacz 2003). The 2D tumor image is padded to produce a square image. Image values are normalized by subtracting the mean value and then dividing by the standard deviation. Pixels outside the tumor are masked to a value of 0. The MATLAB \texttt{fft2} function then produces the Fourier transform of the image. The power spectrum is obtained by taking the square of the modulus of the Fourier transform. The power spectrum is rotationally averaged and plotted on a log-log scale with distance normalized to the full image size. Finally, the slope of a linear fit of this plot summarizes the change in image intensity variation with increasing frequency/decreasing scale (the zero frequency/dc signal is excluded). For the patient images, the slope was calculated for each of the three tumor slices and then averaged.

2.4.2. Fractal analysis. Fractal analysis, like SHARP, is multiscale, and has been widely applied to analyze the structure of tumors and normal anatomic structures. We implemented a fractal analysis method that was recently used to examine MRI of breast tumors (Di Giovanni et al 2012). It was applied to the same three 2D tumor slices that were analyzed using our SHARP algorithm. Each image was downsampled by factors of two; four levels were used (full size, \(\frac{1}{2}\) size, \(\frac{1}{4}\) size, \(\frac{1}{8}\) size). The variation of contrast \((C_v)\) at each spatial resolution was the standard deviation of the pixels of the resampled matrix. Then, a linear fit of \(\ln(\# \text{ voxels})\) versus \(\ln(C_v)\) was performed. Fractal dimension and fractal fit were calculated from this linear
regression. Fractal dimension and fit were then averaged across the three slices to produce per-tumor summary measures.

2.4.3. Texture features. Another widely used texture analysis method is extraction of Haralick features from the gray-level co-occurrence matrix (Yang and Knopp 2011). We implemented this method as described in a paper differentiating benign from malignant breast masses using MRI (Gibbs and Turnbull 2003). The three Haralick features found by Gibbs to be significant predictors of malignancy were used: f4 (variance), f8 (sum entropy), and f9 (entropy). Step size was 1 pixel and 32 gray levels were used. Co-occurrence matrices were calculated for directions of 0°, 45°, 90°, and 135°, then averaged. Parameters were calculated for each of the three tumor slices and then averaged.

3. Results

3.1. Validation studies

We created three classes of synthetic images, meant to approximate tumors with varying appearance, to test whether the SHARP algorithm correctly summarizes the scale of heterogeneity. We randomly generated 20 images of each class and ran the algorithm on each image. Results are presented in figure 4. The algorithm successfully differentiated between the different classes of images. It correctly identified the scale of maximum heterogeneity for the ‘Large blocks’ and ‘Small blocks’ cases, and captured the multiscale heterogeneity of the ‘Large + small blocks’ case.

The three comparison methods were also tested on the synthetic images. Findings are shown in table 1. As image features become more small-scale (‘Small blocks’ image), the power spectrum becomes flatter, indicating greater signal present at higher frequency levels. Fractal dimension scales inversely with the scale of the image features, which matches intuition as the patterns in the ‘Small blocks’ image are more space-filling than those in the ‘Large blocks’ image. Some of the texture results also make intuitive sense. Most adjacent pixels in the ‘Large blocks’ image have similar values, in contrast to the ‘Small blocks’ image, so the f9 entropy value increases with smaller image feature size.

Robustness of the algorithm to high frequency image noise or coil/magnetic field inhomogeneity was tested by adding Gaussian white noise or a smoothly varying intensity gradient across the synthetic images (figure 5). 20 images of each class were generated and then

| Table 1. Comparison algorithm results on synthetic images. Mean value over 20 trials. |
|---------------------------------|----------------|----------------|----------------|
| Method                          | Large blocks  | Large + small blocks | Small blocks   |
| Spectral analysis (slope of power spectrum) | −3.00         | −2.04          | −0.23          |
| Fractal analysis                |               |                |                |
| Fractal dimension               | 1.00          | 1.05           | 1.23           |
| Fractal fit                     | 4.33          | 0.42           | 2.53           |
| Haralick texture features       |               |                |                |
| $f_4$: variance                 | 314.69        | 280.10         | 316.63         |
| $f_8$: sum entropy              | 3.26          | 3.61           | 3.60           |
| $f_9$: entropy                  | 3.52          | 5.70           | 6.42           |
| SHARP $Hetslope_2$              | 0.95          | 0.07           | −0.70          |
| SHARP $MaxScale_2$              | 0.00          | −0.95          | −5.00          |
Figure 5. SHARP output on synthetic images with various perturbations added. The graphs show mean EMD at each scale, averaged across 20 randomly generated images of each class (error bars indicate standard deviation). The added white noise did not significantly affect the heterogeneity versus scale curves. The added linear intensity gradient/bias field mainly affected the Small blocks images at the largest distance scale. When tumor shape was varied (lower panel), output was quite stable.
5%/10% noise or 10/20% gradient were added. With increasing amounts of white noise, the amount of high-frequency / small scale heterogeneity increased slightly for the ‘Large blocks’ and ‘Large + small blocks’ images. Added noise did not affect the plot of heterogeneity versus scale for the ‘Small blocks’ image, likely because this image already consisted of mainly high-frequency noise. Conversely, adding an intensity gradient affected mainly the output for the Small Blocks images, presumably because this was a low-frequency/large scale perturbation being added to an image that started with mostly high frequency signal.

To test the dependence of the algorithm results on the overall shape of the tumor, we created three tumor outlines with various complexity (figure 6). All three shapes were calibrated to have the same area, which allowed the effects of tumor shape to be tested in isolation. The three synthetic images (Large blocks, etc.) were then masked with the three tumor shapes in turn and the SHARP algorithm was run on 20 randomly generated images of each class. Figure 5 shows that the results were quite stable across tumor shapes.

We quantitatively assessed the effects of all the image perturbations on the SHARP algorithm output using intraclass correlation. For each distance scale, this measure compared (1) the variation in EMD between each of the three original synthetic images and the corresponding perturbed image (within-class comparison) and (2) the variation in EMD between synthetic images of the three different classes (between class comparison). The intraclass correlation values for each of the six distance scales were averaged to create a summary measure (table 2). For all amounts of added noise or bias field, the intraclass correlation was >0.8, which indicates good robustness of the results to the perturbations. Similarly, the intraclass correlation between the three tumor shapes was good at 0.89.

### Table 2. Intraclass correlation between original synthetic images and perturbed images. 1 = perfect agreement; 0 = poor agreement.

| Compared images                                      | Mean intraclass correlation (95% confidence interval) |
|------------------------------------------------------|-------------------------------------------------------|
| Original image / 5% Gaussian noise                   | 0.90 (0.84–0.94)                                      |
| Original image / 10% Gaussian noise                  | 0.88 (0.81–0.93)                                      |
| Original image / 10% bias field                      | 0.89 (0.83–0.93)                                      |
| Original image / 20% bias field                      | 0.86 (0.78–0.91)                                      |
| Three tumor shapes (0, 3, 10 lobes)                  | 0.89 (0.85–0.93)                                      |

3.2. Prediction of rhabdomyosarcoma recurrence

We applied the SHARP algorithm to examine the influence of scale of rhabdomyosarcoma tumor heterogeneity on recurrence patterns. Patient characteristics are listed in table 3. Median tumor diameter was 39 mm (range 21–120). The primary tumors tended to harbor
more coarse-scale than fine-scale heterogeneity on contrast-enhanced MRI: scale of maximum heterogeneity was most commonly the full tumor diameter \((n = 11\) using powers of 2 bin spacing, \(n = 7\) using powers of \(\sqrt{2}\) bin spacing). There were 10 disease recurrences, and no deaths without prior disease progression. Using T1-weighted gadolinium-enhanced MRI, the SHARP algorithm produced four heterogeneity measures per patient, which were compared to RFS using univariate Cox proportional hazards analysis. One of the four measures, \(\text{MaxScale}_{\sqrt{2}}\), had a statistically significant association with RFS: coarser scale of maximum SI heterogeneity was predictive of shorter RFS (Cox model log-rank \(p = 0.05\), concordance index 0.71). As seen in figure 7, patients with \(\text{MaxScale}_{\sqrt{2}} > \text{median}\) \((n = 7)\) had median RFS of 15.5 months, versus not reached when \(\text{MaxScale}_{\sqrt{2}} \leq \text{median}\) \((n = 11)\), log-rank \(p = 0.02\). Figure 8 shows examples of tumors with varying scales of heterogeneity.

The other three heterogeneity measures were not predictive of RFS, with Cox model log-rank \(p = 0.24–0.45\). Similarly, Children’s Oncology Group (COG) risk group was not predictive of RFS \((p = 0.35)\) (Malempati and Hawkins 2012). Tumor diameter was not predictive of RFS \((p = 0.19)\).

### 3.3. Comparison with existing methods

In order to compare our SHARP method to existing spatial analysis methods, we also performed spectral, fractal, and texture analyses. Slope of the Fourier power spectrum was not predictive of RFS (Cox model \(p = 0.60\)). Neither fractal dimension nor fractal fit was predictive.

---

**Table 3. Patient characteristics.**

| Category                        | Number of patients |
|--------------------------------|--------------------|
| Age 0–9 years                   | 11                 |
| Age 10–19 years                 | 5                  |
| Age ≥20 years                   | 2                  |
| Sex Female                      | 11                 |
| Sex Male                        | 7                  |
| Histology Alveolar              | 9                  |
| Histology Embryonal             | 8                  |
| Histology Other                 | 1                  |
| Site Parameningeal              | 6                  |
| Site Orbit                      | 5                  |
| Site Genitourinary              | 4                  |
| Site Head and neck              | 2                  |
| Site Extremity                  | 1                  |
| Children’s Oncology Group Low  | 6                  |
| Children’s Oncology Group High | 2                  |
| Group IV                        | 16                 |
| Group III                       | 10                 |
| Stage 1                         | 5                  |
| Stage 2                         | 4                  |
| Stage 3                         | 7                  |
| Stage 4                         | 2                  |
| Initial treatment Chemotherapy  | 18                 |
| Initial treatment Radiation     | 16                 |
| Initial treatment Primary tumor | 5                  |

\(^a\) Malempati and Hawkins (2012).
of RFS ($p = 0.34$, $p = 0.36$, respectively). Similarly, the three Haralick texture features were not predictive ($p = 0.09–0.90$).

4. Discussion

The SHARP algorithm produces a biologically motivated segmentation of tumor regions and reports the amount of heterogeneity at various distance scales. In this study we showed that the algorithm correctly identifies the scale of heterogeneity on synthetic images, is robust to image artifacts and various tumor shapes, and can be applied to routine clinical MRI of rhabdomyosarcoma tumors. In our dataset of 18 patients with rhabdomyosarcoma, RFS was shorter when the primary tumor had coarser scale of heterogeneity on contrast-enhanced MRI. If validated using a larger dataset, this imaging biomarker could be useful for prognostication and to guide allocation to varied treatment intensity. In clinical practice, recurrence risk is currently estimated by the COG risk grouping, which incorporates histology (embryonal versus alveolar), clinical stage, disease site, and extent of resection (Malempati and Hawkins 2012). Treatment intensity is tailored to risk group: treatment algorithms for low-risk patients focus on minimizing side effects, while intermediate- and high-risk patients receive intensive combination chemotherapy. Most of the patients in our study had COG intermediate-risk disease. Outcomes for this risk group have plateaued, with 4-year failure-free survival of 73% in the standard therapy arm of a recent multicenter study (Arndt et al. 2009). Imaging biomarkers such as our heterogeneity measures could help improve prognostication. Other promising markers are being explored, such as tumor \[^{18}\text{F}]\text{fluorodeoxyglucose} (FDG) uptake on post-treatment positron emission tomography (Dharmarajan et al. 2012). New classification schemas could allow for treatment intensification for those patients likely to recur after standard therapy.

The novelty of our method lies in the application of tumor segmentation to analysis of heterogeneity. With the emerging evidence that tumors can be composed of regions with distinct molecular properties, this may be a more biologically plausible approach than other methods.
such as texture analysis. Many texture analysis methods only compare each voxel to its neighboring voxels, reducing the ability to detect larger-scale patterns (Yang and Knopp 2011). Subjective descriptors such as rim enhancement/central necrosis indicate coarse-scale variability of contrast uptake; human-graded enhancement pattern has been found to be predictive of treatment response in breast cancer (Esserman et al 2001). Fractal analysis methods do incorporate multiscale heterogeneity into measures like the fractal dimension (Di Giovanni et al 2012). However, our method relies on different assumptions than fractal methods—specifically, it does not assume self-similarity of the tumor. The complete segmentation produced by the SHARP algorithm also makes it possible to extract a diverse array of features, including the amount of heterogeneity at various distance scales, the area of the tumor with the most/least heterogeneity at each scale, and for a given point in the tumor, the amount of heterogeneity at different scales.

While one of the four SHARP features was associated with rhabdomyosarcoma recurrence, the comparison methods were not predictive of recurrence (Fourier spectral analysis, fractal analysis, Haralick texture features). Edge effects at the border of the tumor are one potential
reason for these negative results. The texture analysis may have been limited by the use of only neighboring pixels for the gray-level co-occurrence matrix. It is also possible the SHARP findings were a false positive result.

This work has several limitations. The constants $C_l$ and $C_X$ used to shape the partition dividing lines were chosen based on visual assessment of segmentations of one patient’s tumor, rather than inherent biological properties of the tumors. The rhabdomyosarcoma dataset is a small sample due to the rarity of this disease, and patient characteristics and MRI parameters were heterogeneous. The small sample size likely explains why COG risk group was not predictive of RFS. Scale of maximum heterogeneity was only a significant predictor of relapse when the smaller bin size was used. This may be because the heterogeneity in the images was consistently greatest at the few coarsest distance scales, which were clumped together by the larger scale bin size. While the rhabdomyosarcomas in this study tended to exhibit mostly large-scale heterogeneity, application of the SHARP method to tumors with more small-scale heterogeneity would need to be performed carefully. This caveat is due to potential of image resolution and partial volume averaging between thick MR slices affecting the amount of heterogeneity seen at small scales. Also, the algorithm cannot be applied to very small tumors: because of the limited resolution of clinical MRI, patients with tumors smaller than 2 cm were excluded from the rhabdomyosarcoma analysis.

We plan to validate our preliminary findings of an association between scale of heterogeneity and disease recurrence using a larger dataset of rhabdomyosarcoma patients and will study the application of the SHARP algorithm to 3D images from breast MRI. Finally, the algorithm could be used to explore change in heterogeneity measures after therapy for assessment of treatment response.

Acknowledgments

We thank Arno Klein, Colin Studholme, and Sharmishtaa Seshamani for helpful discussions.

References

Arndt C A et al 2009 Vincristine, actinomycin, and cyclophosphamide compared with vincristine, actinomycin, and cyclophosphamide alternating with vincristine, topotecan, and cyclophosphamide for intermediate-risk rhabdomyosarcoma: children’s oncology group study D9803 J. Clin. Oncol. 27 5182–8

Agarwal S 2002 Spectral Clustering Toolbox http://homes.cs.washington.edu/~sagarwal/code.html

Ahmed A, Gibbs P, Pickles M and Turnbull L 2013 Texture analysis in assessment and prediction of chemotherapy response in breast cancer J. Magn. Reson. Imaging 38 89–101

Balboa R M and Grzywacz NM 2003 Power spectra and distribution of contrasts of natural images from different habitats Vision Res. 43 2527–37

Brown F M and Fletcher C D 2000 Problems in grading soft tissue sarcomas Am. J. Clin. Pathol. 114 S82–9

Dharmarajan K V, Wexler L H, Gavane S, Fox J J, Schoder H, Tom A K, Price A N, Meyers P A and Wolden S L 2012 Positron emission tomography (PET) evaluation after initial chemotherapy and radiation therapy predicts local control in rhabdomyosarcoma Int. J. Radiat. Oncol. Biol. Phys. 84 996–1002

Di Giovanni F, Ahearn T S, Semple S I, Lovell I M, Miller I, Gilbert F J, Redpath T W, Heys S D and Staff R T 2012 The biological correlates of macroscopic breast tumour structure measured using fractal analysis in patients undergoing neoadjuvant chemotherapy Breast Cancer Res. Treat. 133 1199–206

Esserman L et al 2001 MRI phenotype is associated with response to doxorubicin and cyclophosphamide neoadjuvant chemotherapy in stage III breast cancer Ann. Surg. Oncol. 8 549–59
Fowlkes C, Belongie S, Chung F and Malik J 2004 Spectral grouping using the Nystrom method IEEE Trans. Pattern Anal. Mach. Intell. 26 214–25
Gatenby R A, Grove O and Gillies R J 2011 Quantitative imaging in cancer evolution and ecology Radiology 269 8–15
Gibbs P and Turnbull L W 2003 Textural analysis of contrast-enhanced MR images of the breast Magn. Reson. Med. 50 92–8
Kelly P J, Daumas-Duport C, Kispert D B, Kall B A, Scheithauer B W and Illig JJ 1987 Imaging-based stereotaxic serial biopsies in untreated intracranial glial neoplasms J. Neurosurg. 66 865–74
Malempati S and Hawkins D S 2012 Rhabdomyosarcoma: review of the Children’s Oncology Group (COG) Soft-Tissue Sarcoma Committee experience and rationale for current COG studies Pediatr. Blood Cancer 15 5–10
Mosher T J, Zhang Z, Reddy R, Boudhar S, Milestone B N, Morrison W B, Kwoh C K, Eckstein F, Witschey W R and Borthakur A 2011 Knee articular cartilage damage in osteoarthritis: analysis of MR image biomarker reproducibility in ACRIN-PA 4001 multicenter trial Radiology 258 832–42
Pieper S, Halle M and Kikinis R 2004 3D SLICER Proc. IEEE Int. Symp. Biomed Imaging (Washington, DC) pp 632–5
R Core Team 2013 R: A Language and Environment for Statistical Computing (Vienna: R Foundation for Statistical Computing)
Raunig D L et al (QIBA Technical Performance Working Group) 2015 Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment Stat. Methods Med. Res. at press
Rubner Y, Tomasi C and Guibas L 2000 The earth mover’s distance as a metric for image retrieval Int. J. Comput. Vision 40 99–121
Shi J and Malik J 2000 Normalized cuts and image segmentation IEEE Trans. Pattern Anal. Mach. Intell. 22 888–905
Snuderl M et al 2011 Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma Cancer Cell 13 810–7
Yang X and Knopp MV 2011 Quantifying tumor vascular heterogeneity with dynamic contrast-enhanced magnetic resonance imaging: a review J. Biomed. Biotechnol. 732848