Imaging Intratumor Heterogeneity: Role in Therapy Response, Resistance, and Clinical Outcome

James P.B. O’Connor1,2, Chris J. Rose1, John C. Waterton1,3, Richard A.D. Carano4, Geoff J.M. Parker1, and Alan Jackson1

Abstract

Tumors exhibit genomic and phenotypic heterogeneity, which has prognostic significance and may influence response to therapy. Imaging can quantify the spatial variation in architecture and function of individual tumors through quantifying basic biophysical parameters such as CT density or MRI signal relaxation rate; through measurements of blood flow, hypoxia, metabolism, cell death, and other phenotypic features; and through mapping the spatial distribution of biochemical pathways and cell signaling networks using PET, MRI, and other emerging molecular imaging techniques. These methods can establish whether one tumor is more or less heterogeneous than another and can identify subregions with differing biology. In this article, we review the image analysis methods currently used to quantify spatial heterogeneity within tumors. We discuss how analysis of intratumor heterogeneity can provide benefit over more simple biomarkers such as tumor size and average function. We consider how imaging methods can be integrated with genomic and pathology data, instead of being developed in isolation. Finally, we identify the challenges that must be overcome before measurements of intratumoral heterogeneity can be used routinely to guide patient care.

Introduction

Malignant tumors are biologically complex and exhibit substantial spatial variation in gene expression, biochemistry, histopathology, and macroscopic structure. Cancerous cells not only undergo clonal evolution from a single progenitor cell into more aggressive and therapy-resistant cells, but also exhibit branched evolution, whereby each tumor develops and preserves multiple distinct subclonal populations (1). This genetic heterogeneity (1–2), combined with spatial variation in environmental stressors, leads to regional differences in stromal architecture (3), oxygen consumption (4, 5), glucose metabolism (4), and growth factor expression (6). Consequently, tumor subregions develop, each with spatially distinct patterns of blood flow (7, 8), vessel permeability (9), cell proliferation (10), cell death (11), and other features.

Spatial heterogeneity is found between different tumors within individual patients (intradrug tumor heterogeneity) and within each lesion in an individual (intra-tumor heterogeneity). Intradrug heterogeneity is near ubiquitous in malignant tumors, but the extent varies between preclinical cancer models and between patients (12). Allowing for these differences, some common themes emerge. First, intratumor heterogeneity can be dynamic. For example, variations in tumor pO2 fluctuate over minutes to hours (5, 6). Second, intratumor heterogeneity tends to increase as tumors grow (7, 13). Third, established spatial heterogeneity frequently indicates poor clinical prognosis (14), in part due to resistant subpopulations of cells driving resistance to therapy (3, 15). Finally, intratumor heterogeneity may increase or decrease following efficacious anticancer therapy (11, 16), depending on the imaging test used and the underlying tumor biology (17).

Imaging depicts spatial heterogeneity in tumors. However, while imaging is central to diagnosis, staging, response assessment, and recurrence detection in routine oncologic practice, most clinical radiology and research studies only measure tumor size or average parameter values, such as median blood flow (18). In doing so, spatially rich information is discarded. There has been considerable effort to use more sophisticated analyses to either quantify overall tumor spatial complexity or identify the tumor subregions that may drive disease transformation, progression, and drug resistance (11, 19).

In this review, we highlight the strengths and weaknesses of methods that measure intratumor spatial heterogeneity (Fig. 1 and Table 1). We evaluate evidence that heterogeneity analyses provide any clinical benefit over simple “average value” measurements. We discuss how imaging, genomic, and pathology biomarkers of intratumor heterogeneity relate to one another. Finally, we identify the hurdles to translating image biomarkers of spatial heterogeneity into clinical practice.

Qualitative Assessment of Heterogeneity

Radiologists use qualitative descriptors to describe adverse spatial features and functional heterogeneity on clinical scans.
O’Connor et al.

For example, when assessing pulmonary nodules on CT (20) and breast lumps on X-ray mammography (21), spiculation implies greater risk of malignancy compared with well-circumscribed lesions. Indeed, spiculate morphology is part of the BI-RADS (Breast Imaging Reporting and Data System) lexicon that classifies breast lesions as “radiologically malignant” (22).

Identifying a tumor “hot spot” is also commonplace in cancer radiology. The maximum standardized uptake value (SUV_{max}) derived from 18F-FDG PET-CT imaging is an established proxy for identifying abnormal glucose metabolism, based on identifying the one or more voxels with greatest abnormality. Measuring SUV_{max} is simple and reproducible (23, 24) and can be performed on clinical work stations. The presence of abnormal glycolysis in part of a tumor using SUV_{max} is used widely to stage and monitor response in several malignant tumors (25). In glioma, regional high values of tracer uptake (18F-FDG and 11C-methionine) have been used to grade tumors using targeted biopsy (26).

Perfusion CT and MRI methods use “hot spot” analysis to identify tumor regions with the most abnormal vascular features. Specialist neuro-oncology centers use dynamic contrast-enhanced MRI (DCE-MRI) or dynamic susceptibility contrast MRI (DSC-MRI) in patients with high-grade glioma (HGG) to map relative cerebral blood volume (rCBV; ref. 27) based on the rationale that “more vascular” regions correspond to highest malignant grade and that this improves prognostic assessment in patients. It is important to note that hot spot analyses are subjective, only identify regions that are maximum (or minimum when gray scale is inverted), and that observer evaluation of heterogeneity is highly influenced by display ranges and color schemes (28) unless objective algorithms are used.

**Figure 1.**

Quantifying intratumoral heterogeneity: the example liver metastasis from a patient with a colonic primary tumor can be measured in several different ways. A, most clinical assessment of tumors is size based. B, functional imaging methods can measure tumor pathophysiology but tend to derive average parameter values, such as median K^trans. C, some intratumoral heterogeneity methods quantify overall complexity of a distribution (histograms) or spatial arrangement of data (texture analysis). Other methods identify tumor subregions using a priori assumptions (partitioning) or data-driven approaches (multispectral analysis).

**Voxels: Considerations for Quantifying Heterogeneity**

Imaging modalities measure biophysical signals in tissues and spatially encode these signals to create clinical images or parameter maps composed of three-dimensional picture elements called voxels. Heterogeneity analyses combine data from many voxels. Several issues arise when interpreting these data.

First, some voxels suffer from partial volume averaging (typically at interface with nontumor tissue). Second, there is inevitable compromise between having sufficient numbers of voxels to perform the analysis versus sufficiently large voxels to overcome noise and keep imaging times practical (29). Most methods of analysis require hundreds to thousands of voxels for robust application. Third, many studies of tumor spatial heterogeneity have used standard clinical data from protocols dictated by clinical rather than research needs. In some cases, sections of tumor were omitted when noncontiguous tumor sampling was used (30), which may confound three-dimensional spatial analyses (31). Fourth, some calculated voxel values, such as apparent diffusion coefficient (ADC), contrast transfer coefficient (K^trans), and blood flow, are derived from multiple images obtained over time. The estimation errors associated with motion vary for different parameters and for different voxels, which should be considered when assessing tumor heterogeneity.

Finally, CT, MRI, or PET voxels are usually nonisotropic (slice thickness exceeds in-plane resolution). Dimensions are typically 200 to 2,000 μm for rodent models and 750 to 5,000 μm for clinical tumors. Compared with genomic and histopathology biomarkers, this represents many orders of magnitude difference in scale (32).
making it difficult to validate image heterogeneity biomarkers against pathology. However, imaging methods have the distinct advantage of whole-tumor sampling, allowing all of the genetic and pathologic variation within tumors to be sampled (33).

### Measuring Degree of Heterogeneity: Quantifying Parameter Distributions

Voxel values can be plotted as histograms, from which many simple descriptors can be extracted as potential biomarkers. These include simple descriptors of image heterogeneity such as standard deviation, interquartile range, nth centile(s), skew and kurtosis (Supplementary Fig. S1), as well as mean and median values (17, 34). In these approaches, the inherent spatial relationship between voxels is discarded and data are treated as a list of continuous variables.

#### Histogram analysis

Histograms can be generated using widely available software and have proved popular methods for characterizing intratumoral heterogeneity, accounting for approximately half of all published studies (35). Several important points should be considered. Histogram analyses have high dimensionality and generate many parameters and thus require correction for multiple comparisons (36). The repeatability and reproducibility of many histogram-derived parameters are uncertain and have not yet been evaluated in multicenter studies (29). Furthermore, many parameters, such as fifth centile or kurtosis, have no clear biologic correlate, making biological validation difficult.

#### Evidence for clinical benefit

More than 200 histogram-based studies have analyzed imaging data of tumor response or outcome (PubMed search performed on October 22, 2014), with rapid rise in numbers recently (Supplementary Fig. S2). Unfortunately, many of these studies were retrospective, performed in small numbers of patients, and did not demonstrate added benefit over simpler measurements of tumor structure and function. For example, in HGG, various histogram parameters have been correlated with overall survival (OS; \( P < 0.05 \)), but relationships were generally equivalent or weaker than those seen with median rCBV values (37). Other studies have also shown marginal benefits. Measurements such as peak height of the CBV histogram had superior sensitivity and negative predictive value over hot spot analysis for distinguishing low-grade glioma from HGG (histogram, 90% sensitivity; hot spot, 55%–76% sensitivity; ref. 38).

The real value in histogram analysis appears two-fold. First, changed heterogeneity in the data distribution may relate to clinical outcome. In cervical cancer, more heterogeneous distributions of FDG-PET voxel values were related to greater risk of lymph node metastases, risk of local recurrence, and worse progression-free survival (PFS; ref. 39). Preclinical data have shown that regional response to therapy may result in unimodal histograms becoming bimodal (ref. 41; Fig. 2).

Second, histograms can quantify data with complex distributions, where average values may mask important information within the data. For example, in patients with HGG receiving bevacizumab and cytotoxic therapies, pretreatment ADC values had a bimodal distribution. Overall mean ADC did not relate to subsequent PFS (\( P = 0.14 \)), but the mean ADC value of the lower mode related significantly to PFS (\( P = 0.004 \); ref. 30), suggesting that MRI identified two distinct tumor subregions (Supplementary Fig. S1). These findings have been replicated in a multicenter study.
suggesting possible clinical translation although when histograms from different tumors differ in both location and degree of stretch along the $x$ axis, this can hinder analysis of cohorts of histogram data. Attempts to model for distribution shape may address this problem (43, 44), but remain in their infancy.

**Measuring Degree of Heterogeneity: Quantifying Spatial Complexity**

Feature analyses measure the spatial complexity of objects. Unlike histogram analysis, the spatial arrangement of voxel values is retained. Related methods, including texture analysis, fractal techniques, and Minkowski functionals, have been applied to tumor data and account for approximately half of studies measuring tumor heterogeneity (35). Initial reports suggest that feature-based metrics derived from $^{18}$F-FDG PET (45), $^{18}$F-FLT PET (46), CT (47), and MRI (48) have good limits of agreement, with coefficients of variation comparable with summary statistics derived from the same data.

**Texture analysis**

There is a large body of literature concerning texture analysis in breast cancer. Many studies use the Haralick method (49), where a co-occurrence matrix element denoted $P_{i,j}$ measures the probability of starting from any image voxel with designated value $i$, moving $d$ voxels along the image in direction $\theta$, and then arriving at another voxel with value $j$. This co-occurrence matrix is a two-dimensional histogram describing a joint distribution of all the possible moves with step size $d$ and direction $\theta$ on the image. From this, various extracted features, including lesion contrast and correlation, can estimate shape and/or spatial complexity.

**Fractal analysis and Minkowski functionals**

Fractal dimensions estimate the complexity of geometric patterns resulting from abstract recursive procedures (50). The simplest fractal dimension is the box-counting dimension ($d_0$), computed by imposing regular grids of a range of scales on a binary object in question and then counting the number of grid elements (boxes) that are occupied by the object at each scale (Supplementary Fig. S3). The box-counting dimension is the slope of the line of best fit when plotting the number of occupied boxes against the reciprocal of the scale on log–log axes (51). Increasingly complex variants can incorporate continuous scale values of parameters such as Hounsfield unit density or $K_{trans}$. Minkowski functionals analyze binarized images over a range of thresholds and also quantify space-filling properties of tumors.

**Evidence for clinical benefit**

Texture analyses have been used extensively in X-ray mammography (52). Early applications included discrimination of glandular and fatty regions in mammograms (53) and distinguishing benign and malignant lesions (31). Minkowski functionals analysis can improve stratification of risk for developing breast cancer (receiver operating curve area under curve of $>0.9$; ref. 54). Here, more heterogeneous images indicated more aggressive tumors. Feature analyses show the greatest promise as prognostic indicators. CT-based feature analysis parameters predicted OS independent of tumor stage in patients with colorectal cancer treated with cytotoxic therapy ($P < 0.01$; ref. 55), predicted time to progression in patients with renal cancer treated with various antivascular therapies ($P = 0.005$; ref. 56) and predicted OS in non–small cell lung cancer ($P = 0.046$; ref. 57). Similarly, MRI fractal biomarkers have shown prognostic relationships in patients with colorectal cancer treated with bevacizumab.
(P < 0.00005; ref. 58) and in patients with sarcoma treated with cytotoxic therapy (ref. 59; P < 0.005). These approaches are now being explored and validated in large populations using an approach termed “radiomics,” through which existing clinical imaging data are mined to identify heterogeneity features that predict clinical outcome (60).

**Identifying Tumor Subregions**

Tumor images contain hundreds to thousands of voxels. Grouping "similar" voxels together (parcellation) may define multiple subregions with common biology that respond differentially to therapy or drive progression. Parcellation techniques differ in their underlying assumptions and methodology. Some use prior knowledge, whereas others rely solely on information contained within the images (Fig. 3).

**Parcellation using a priori assumptions**

Voxels can be categorized by presence or absence of a hallmark, such as enhancement. Alternatively, tumor signals (such as CT density, 18F-FDG SUVmax, or $K^{trans}$) can be parcellated using one or more thresholds. In distinction, tumor regions may be defined geographically, for example by modeling tumors as spheres with concentric radial subregions (61) or by labeling voxels as “rim” or “core” based on relative voxel position in histogram distributions (62).

These simple approaches have important caveats. "Binary" features such as enhancement are not absolute but depend on how images are acquired and analyzed. Threshold values for continuous data based on "cut points" selected to enhance statistical separation in studies are arbitrary. A priori methods may have a misleading biologic basis. The Macdonald criteria in HGG illustrate this point, as measuring contrast-enhancing tumor only has been superseded by the Response Assessment in Neuro-Oncology criteria, which incorporate measures of nonenhancing tissue (with likely microscopic foci of infiltrative neoplastic cells) into response criteria (63).

**Data-driven segmentation**

Some studies acquire multiple imaging parameters (for example, $K^{trans}$, ADC, and 18F-FDG SUVmax) and analyze each signal independently (64). An alternative strategy parcellates voxels with similar signals (or "spectra") into functionally coherent regions within a lesion (65). Most multispectral analyses use pattern recognition techniques that simultaneously analyze images to identify voxel clusters in a multidimensional feature space. A classifier then groups individual voxels together based on their similarities and differences (66).

**Evidence for clinical benefit**

Multiple studies of patients with HGG have shown lower baseline lesion enhancing tumor volume (ETV) had beneficial OS ($P = 0.0026$; ref. 67) or PFS ($P = 0.0309$; ref. 68) and that early reduction in ETV after bevacizumab related to OS ($P = 0.0008$) (69), whereas WTV and $K^{trans}$ did not. Studies of solid tumors outside the brain have shown that ETV or the proportion of enhancing voxels ($E_T$) before treatment is correlated with PFS in cervical cancer (70, 71). Trials of antiangiogenic drugs suggest that $E_T$ provides pharmacodynamic information, independent from other DCE-MRI parameters (48, 72). These data support the hypothesis that measuring tumor regions may be a more useful biomarker than average values of whole tumors in some clinical scenarios.

Threshold-derived partitioning has shown value in patients with gastrointestinal stromal tumors (GIST) imaged with

---

**Figure 3.**

Quantifying tumor subregions: a priori methods parcellate tumor regions based on binary classifiers [e.g., nonenhancing (−) or enhancing (+) voxels from a $K^{trans}$ map] or threshold classification using a “cut point” defined from previous data, or an arbitrary value such as the median (e.g., of an ADC distribution). Multispectral classification generates clusters based on data-driven segmentation from multiple signals (e.g., $K^{trans}$ and ADC), for example using principal components (PC) analysis (graph shows three clusters based on PC1 and PC2). The volume or proportional fraction of each region is measured in these methods. Geographic methods are distinct because they define subregions based on the location of a pixel rather than its parameter value(s). BM, biomarker.
Here, tumor SUV below 8 g/mL 4 weeks after treatment with sunitinib was associated with markedly longer PFS than those with SUV above the same level (29 vs. 4 weeks; \( P < 0.0001 \); ref. 73). Similarly, SUV thresholds have distinguished good and poor time to treatment failure in GIST patients treated with imatinib (\( P < 0.0001 \); ref. 74). Furthermore, there is some evidence that the presence of high SUVmax values may relate to OS in some cancers (75). In these studies, cutoff points have been chosen using \textit{post hoc} criteria. Prospective studies are required to validate these thresholds in specific patient–therapy combinations if these techniques are to have clinical translation.

Data-driven approaches have successfully distinguished viable tumor from nonviable tumor using multiparametric MRI and validated the method against H&E histology (76, 77). Moreover, image-defined regions of viable and nonviable tumor show differential response to radiotherapy (78) and to anti-VEGF antibodies (ref. 79; Fig. 4). Multispectral analyses of baseline data that identify tumor subregions with distinct biology, responsible for driving tumor response, resistance, and progression, are highly attractive, although performed rarely. Algorithms to use for this type of analysis are commercially available (80).

Unfortunately, it is very difficult to track change in individual voxels. Some progress has been made in studies of HGG using a method termed "response parametric mapping," in which rCBV and ADC maps have been categorized as showing no change or increasing/decreasing by greater than 20% following therapy. The proportion of tumor with reduced rCBV was correlated with OS (\( P = 0.019 \)), where mean rCBV was not (81). However, it is extremely difficult to extend this approach to accommodate changes in tumor size, orientation, and deformation, making it difficult to translate these methods from specialist laboratories to health care systems.

**Integrating Imaging, Genomics, and Histopathology**

Recent work highlights the intratumor variation in gene mutation and expression (33), but few studies have explored the spatial relationship among imaging, genomics, and histopathology. Preclinical studies have reported differential levels of gene activation and protein transcription that relate to regional perfusion as measured with DCE-MRI (82) and related differential gene expression to regional PET signals (83). Pilot data have associated heterogeneous enhancement patterns with genetic subtypes of breast cancer (84). The ADC characteristics at tumor margins, rather than overall mean values, have been shown to correctly categorize oligodendrogliomas by their 1p/19q loss status (85).

These data highlight the need for larger prospective studies to elucidate how and when integration of imaging, genomic, and pathology data may be useful. These studies must evaluate large complex data across a range of biologically different scales (15, 32). Two questions require urgent answers. First, do imaging, genomics, and histopathology show spatial correspondence because they \textit{measure the same biology} in different ways, and if so,
can imaging reliably identify important subregions non-invasively? If so, imaging could extend personalized medicine to tumors that cannot be biopsied safely. Alternatively, do imaging, genomics, and histopathology measure different biology, providing complementary data? If so, this would open up new multidisciplinary strategies for advancing personalized medicine focusing phenotype and genome together, rather than genome in isolation.

**Future Directions and Conclusions**

Diverse philosophical and mathematical approaches can quantify intratumor heterogeneity. Most techniques can be applied across all imaging modalities, while some require datasets with multiple signals. Each method has strengths and limitations (Table 2). Advances in hardware such as simultaneous datasets with multiple signals. Each method has strengths and weaknesses and can reliably identify important subregions non-invasively. Imaging can become an area of intense interest, attempting to parallel advances in next-generation sequencing. This "radiomic" approach (60) places heterogeneity parameters at its center, but the sheer number of parameters under investigation incurs statistical dangers due to large numbers of multiple comparisons. As these challenges are addressed, heterogeneity analyses have great potential to translate from the research arena to play a role in clinical decision making.

**Disclosure of Potential Conflicts of Interest**

J.C. Waterton is an employee of and has ownership interest (including patents) in AstraZeneca. R.A. Carano is an employee of Genentech and has ownership interest (including patents) in AstraZeneca. C.J.M. Parker is CEO of and has ownership interest (including patents) in Bioxydyn Limited; reports receiving commercial research grants from AstraZeneca, Merck Serono, and Roche; and is a consultant/advisory board member for GlaxoSmithKline. No potential conflicts of interest were disclosed by the other authors.

**Grant Support**

J.P.B. O'Conor is supported by a Cancer Research UK Clinician Scientist Fellowship (C19221/A15267). J.P.B. O’Connor, G.J.M. Parker, and A. Jackson are supported by a Cancer Research UK and EPSRC Comprehensive Imaging Centre grant (C8742/A018097).

Received May 27, 2014; revised October 22, 2014; accepted October 28, 2014; published OnlineFirst November 24, 2014.

---

**References**


---

**Table 2.** Image analysis methods for quantifying spatial heterogeneity: comparison of strengths and weaknesses

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Availability</th>
<th>Literature</th>
<th>Spatial information</th>
<th>Single-center reproducibility</th>
<th>Potential statistical and design issues</th>
<th>Biologic correlate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative interpretation by radiologist</td>
<td>Worldwide</td>
<td>Extensive</td>
<td>Evaluated qualitatively</td>
<td>Largely unknown</td>
<td>Subjective and difficult to apply in quantitative study</td>
<td>Measure of overall complexity</td>
</tr>
<tr>
<td>Hot spot</td>
<td>Worldwide</td>
<td>Extensive</td>
<td>Ignored outside hot spot</td>
<td>Very good</td>
<td>Subjective</td>
<td>Difficult to evaluate</td>
</tr>
</tbody>
</table>

**Quantitative measurement by scientist**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Availability</th>
<th>Literature</th>
<th>Spatial information</th>
<th>Single-center reproducibility</th>
<th>Potential statistical and design issues</th>
<th>Biologic correlate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram analysis</td>
<td>Extensive</td>
<td>Extensive</td>
<td>Discarded</td>
<td>Largely unknown</td>
<td>Large number of parameters derived: multiple comparisons</td>
<td>Difficult to evaluate</td>
</tr>
<tr>
<td>Feature analysis</td>
<td>Specialist technique</td>
<td>Moderate</td>
<td>Used to compute spatial complexity</td>
<td>Very good</td>
<td>Large number of parameters derived: multiple comparisons</td>
<td>Measure of overall complexity</td>
</tr>
<tr>
<td>A priori segmentation</td>
<td>Specialist technique</td>
<td>Small</td>
<td>Used to define subregions</td>
<td>Largely unknown</td>
<td>Various methods; category definition may be arbitrary</td>
<td>Subregion identified: can match to histology</td>
</tr>
<tr>
<td>Multispectral analysis</td>
<td>Specialist technique</td>
<td>Minimal</td>
<td>Used to define subregions</td>
<td>Unknown</td>
<td>Multidimensional datasets</td>
<td>Subregion identified: can match to histology</td>
</tr>
</tbody>
</table>

---

1In some studies.


25. Knopp EA, Cha S, Johnson G, Mazumdar A, Gol...


