Significance of Image Heterogeneity in Multi-Modality Medical Imaging

Devin Miles, Department of Physics, Case Western Reserve University

A thesis submitted to the Senior Project Committee, Department of Physics, Case Western Reserve University
# Table of Contents

I. Abstract, 3  
II. Introduction, 4  
III. Background, 6  
  a. Positron Emission Topography  
  b. Multi-modal Imaging  
  c. Phantoms  
  d. Biological Heterogeneity  
  e. Intrinsic Imaged Heterogeneity  
IV. Objectives, 10  
V. Literature Review, 11  
VI. Methods, 14  
VII. Results, 17  
VIII. Discussion, 23  
IX. Conclusion, 25  
X. Acknowledgements, 26  
XI. Citations, 27  
XII. Appendix: Code, 28
I. Abstract

Heterogeneous features have been observed in clinical PET-CT and PET-MRI imaging of tumors. The biological significance of this heterogeneity is not well understood but may have prognostic value for clinicians. To advance understanding of this parameter, an investigation of the inherent heterogeneities of PET imaging has been performed using known homogeneous subjects and artificial heterogeneous subjects. This investigation was performed through the use of a previously established metric of heterogeneity quantification. To accurately use this metric, we found it necessary to develop a normalization generalized to image bit depth. We found the heterogeneity factor to hold validity through phantom studies, although a size dependence was observed as a result of small image size, low resolution, and high magnitude of noise.
II. Introduction

Tumor heterogeneity has always been a confounding factor in cancer treatment, due to its effects on metabolic activity, tumor structure, and unpredictability. Its origins can be traced back to trends in cancer growth, which can be thought of as an example of Darwinian evolution- where a tumor begins as a single cell and reproduction introduces various mutations. These mutations can be more or less advantageous to the survival of the cells, where the more resilient live on to grow, reproduce, and mutate further. This process occurs sporadically throughout a tumor, creating an overall irregular, heterogeneous mass of tissue.\(^1\)

![Figure 1: An example heterogeneous tumor, imaged using PET/CT\(^6\)](image)

While this biological heterogeneity is already highly complex and unpredictable, varying degrees of heterogeneity are typically introduced into medical images from the imaging modality itself. Intrinsic image heterogeneity (also referred to as “image inhomogeneity” in medical physics) is variation in the intensity of the image signal that does not correspond to the physical characteristics of the imaged subject. Some work has been performed to characterize image inhomogeneity, which can arise from the imaging method,
attenuation correction, and reconstruction method, among other sources.\cite{2} The role of intrinsic heterogeneity in the heterogeneity of imaged tumors has not been systematically investigated, and the implications of tumor heterogeneity at imaging for medical diagnosis and prognosis are mostly qualitative in current clinical imaging practice.\cite{1} Image heterogeneity could be largely eliminated through the use of correction algorithms, but this runs the risk of destroying the subtle features of the real image.\cite{3}

Previously, tumor heterogeneity has typically been evaluated visually or through microdissection, which falls prey to sampling issues based on the size of the tumor specimen.\cite{4}\cite{1} To make these appraisals less subjective to human error and more meaningfully descriptive of the heterogeneities within the cancer, a quantitative means of describing heterogeneity becomes necessary.
III. Background

a. Positron Emission Tomography (PET)

A radiopharmaceutical, typically fluordeoxyglucose (FDG), labeled with fluorine-18, is injected into a patient or specimen followed by a short period of time allotted by a technician to allow the drug to achieve a desired distribution in the body via physiological processes of perfusion and metabolism. Once properly distributed, the patient is taken to a PET machine and imaging is performed. The radioisotope emits positrons, which, after a short distance, annihilate with electrons in the body. This process emits two antiparallel photons that are detected on a ring of scintillator/photomultiplier tube detectors. Coincident detections of photon pairs are marked and used to reconstruct an image. \(^8\)

While more efficient methods are currently being developed, filtered back projection (FBP) is a common reconstruction algorithm for PET imaging. Although fast and computationally cheap to perform, FBP has a tendency to incorporate shot noise to the resultant image. Also, FBP is unable to account for true randomness from data acquisition, which can often result in lower resolution images.

While CT and MRI are more capable of accurately depicting size, location, and tumor shape, PET imaging has the unique ability to depict metabolic activity within a lesion. Lesions of high metabolic activity tend to be more aggressive, and aggressive tumors tend to be more difficult to treat.

b. Multi-Modal Imaging
A recent trend in the development of new imaging equipment is to create machines capable of imaging more than one modality in a single procedure or session. As a result, PET-CT and, more recently, PET-MRI, are becoming more common tools in many hospitals. For this work, PET-CT and PET-MRI systems from the same manufacturer were available, and the PET components of both systems are very similar in form and function. The main difference lies in the attenuation correction algorithm, which uses CT or MRI image information to correct the PET image reconstruction algorithm for the attenuation of some photons by the body.

In PET-CT, the attenuation correction is based on measurements of tissue density taken from the CT component. While generally useful in improving image quality, the correction can suffer from flaws in the CT data, namely from respiratory patterns, various contrast agents, implants, bias due to beam hardening, or scattering.[13]

In PET-MRI, the attenuation correction cannot be measured directly from the MRI images, because the MRI signal intensity does not correspond to tissue density or PET photon attenuation. Current approaches segment the MR images to separate the data into regions of solid tissue, air (outside the body), and intermediate density (in lung regions). A map is constructed where uniform attenuation values are applied to the three regions. Such maps are less detailed than CT-based attenuation correction, but they can be used to perform acceptable attenuation correction of PET images and achieve quantitative corrected PET image values that agree well with PET-CT imaging in patients and phantoms.[14]
c. Phantoms

Phantoms are tools used to test for image quality without the need for a human patient.[7] Designs can vary depending on the desired body part, patient size, and imaging modality. Within the scope of this experiment, water PET phantoms were chiefly used. They consist of a large acrylic tank with six small reservoirs- a small concentration of FDG is placed in the large compartment of the phantom to simulate background from surrounding tissue and each reservoir is filled with a much higher concentration to simulate regions of interest with higher uptake values.

Figure 2: a water phantom[15]


d. Biological Heterogeneity

It is widely believed that cancer growth begins from a single ‘cancer stem cell.’ As the lone cell grows and multiplies, random mutations can occur that may be more or less advantageous to the survival of this new breed of tissue. Similar to Darwinian evolution, varying populations of cancerous cells with new mutations and biological features can be observed, as well as areas of dead, necrotic, or oxygen-starved tissue depending on the tumor's size and location. As a result of these biological differences, varying regions within a tumor may have higher or
lower rates of glucose uptake, which can result in different 'light and dark' regions within medical images.[1]

e. Intrinsic Imaged Heterogeneity

In addition to the biological heterogeneity that can be observed in most imaged tumors, a layer of heterogeneity may also be imparted onto the image from the imaging modality itself.[8] This inherent inhomogeneity can result from many aspects of the imaging process, some of these being edge effects, reconstruction mechanism, patient movement, random scattering, attenuation correction, hard boundaries, and even pixel size. Generally, very little can be done to eliminate this type of heterogeneity, yet it has the potential to obscure subtle details and tumor structure in severe cases.
IV. Objectives

We hypothesized that the heterogeneity factor would characterize systematic heterogeneity present in medical images due to the imaging method, as well as heterogeneity of specified biological regions of interest (ROI). We hypothesized that this factor could be efficiently calculated for both phantom and clinical studies using PET.

We planned to implement and evaluate an effective means of quantifying image heterogeneity using multi-modal imaging equipment. To test this new metric, we performed imaging experiments on known homogeneous and known heterogeneous materials and, as needed, designed and built phantoms and test objects to support these imaging experiments.
V. Literature Review/Previous Methods

Aside from visual inspection, traditional approaches of describing tumor heterogeneity have involved microdissection and analyses of biological composition. Within this branch of thought, approaches are generally focused on examining specific genetic loci or entire genomes within the cancerous tissue. Analyses of specific loci are useful in the investigation of genes that drive the progression and growth of the tumor, although the typical sample size is usually not indicative of the full lesion and data interpretation is generally very complex. Similarly, analysis of entire genomes within a tumor can reveal similar results, with the caveat that larger genetic trends can be detected but at the cost of resolution and definitive knowledge of function. Both of these approaches, while revealing about the underlying traits of an individual tumor, can suffer from sampling issues and are fairly time-consuming. By the time a full analysis can be performed, more aggressive instances of cancer could further evolve from the original sample, which would no longer be indicative of the living lesion.\[1\]

As a result of these limitations, as well as the prevalence of imaging technology, many groups have made efforts to develop a more convenient means of describing tumor heterogeneity, in particular through quantitative analysis. Over the years, a multitude of quantitative methods have been envisioned and tested, with much variation based around imaging modality and specific traits of focus.
One group aimed to describe heterogeneity through defining a functional risk volume, as opposed to the total anatomical volume typically used.\cite{11} Data prior to, and after, treatment was gathered via dynamic contrast-enhanced MRI. Voxels in the image with low signal intensity were the result of low perfusion of the contrast agent within the lesion and were ignored. Functional risk volume was then calculated from the summation of the remaining voxels corresponding to areas of high perfusion. Results of this method show correlation between high functional risk volume and decreased survival rate.

Other groups have instead chosen to use texture analysis and drug uptake values, as opposed to perfusion, to quantify heterogeneity.\cite{10} Prior data had claimed drug uptake to be less useful than other means, so the aim here was to test for reproducibility of results. Multiple PET images were gathered of a group of subjects over a few days, followed by an array of data manipulations to analyze texture in the images, based around pixel intensity, intensity difference, fraction of high and low intensity

<table>
<thead>
<tr>
<th>Feature</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-area emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{z(i,j)}{\bar{p}^2} )</td>
</tr>
<tr>
<td>Large-area emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} z(i,j) )</td>
</tr>
<tr>
<td>Intensity variability</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{M} \left[ \sum_{j=1}^{N} \frac{z(i,j)}{\bar{p}^2} \right] )</td>
</tr>
<tr>
<td>Size–zone variability</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{z(i,j)}{\bar{p}^2} )</td>
</tr>
<tr>
<td>Zone percentage</td>
<td>( \Omega / \sum_{i=1}^{M} \sum_{j=1}^{N} \bar{p}^2 z(i,j) )</td>
</tr>
<tr>
<td>High-intensity emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{p} z(i,j) )</td>
</tr>
<tr>
<td>Low-intensity small-area emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{z(i,j)}{\bar{p}^2} )</td>
</tr>
<tr>
<td>High-intensity small-area emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{p}^2 z(i,j) )</td>
</tr>
<tr>
<td>Low-intensity large-area emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{p} z(i,j) )</td>
</tr>
<tr>
<td>High-intensity large-area emphasis</td>
<td>( \frac{1}{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{p}^2 z(i,j) )</td>
</tr>
</tbody>
</table>

\( \Omega = \text{number of homogeneous areas within tumor}; z = \text{intensity size–zone matrix}; M = \text{discretization value}; N = \text{size of largest homogeneous area within tumor}; z(i,j) = \text{number of areas with intensity } i \text{ and size } j \).
areas, and many others. Results across multiple images were varied across each form of analysis, but the highest reproducibility was attained through measurements of maximum and average uptake values.\textsuperscript{[9]}

Lastly, an assessment was performed to compare the performance of a qualitative analysis from physicians with textural analysis of FDG uptake, in regards to treatment response and survival rates. Images were taken prior to treatment, and treatment responses were assessed two years after. Both methods of analysis correlated well with patient response to treatment.\textsuperscript{[12]}

With all of this in mind, we intended to utilize a method capable of accounting for areas of varying uptake and perfusion, utilizing texture analysis, and allowing for high correlation with a qualitative image assessment.
VI. Methods

Following the methods developed by Drs. Brooks and Grigsby, a region of interest from a single PET image was isolated and placed onto a uniform, black background of 0 intensity. A simple filter is passed over this image, setting any pixels with an intensity of less than 60% of the maximum intensity to 0, as they are not viewed as a living part of the tumor. For each iteration of this calculation, a line is drawn between a pair of pixels (labeled by the subscripts \(m\) and \(n\)), and each pixel

\[
I(r_{ml}) = I_m + \frac{I_n - I_m}{r_{mn}} r_{ml},
\]

that falls along that line is involved in this step. An intensity difference is calculated between the starting pixel and each pixel \((l)\) along the line.

These intensity differences are summed and normalized to the length of the line to give the average intensity difference between this pixel pair.

\[
\Delta \overline{I} = \frac{1}{L} \sum_{l \in \ell} |I(r_{ml}) - I_l|
\]

This is repeated for each unique pair of pixels within the ROI, and the resultant list of averages are grouped by length and again averaged.

\[
\zeta \equiv \frac{1}{0} \int \Delta \overline{I} \left( \frac{L}{\bar{L}} \right) d\left( \frac{L}{\bar{L}} \right)
\]

The average of averaged intensity differences are integrated by length/(max length) and the numerical result is the heterogeneity factor.
This technique was designed to correlate well with a qualitative assessment from a physician or radiologist. Within their previous work, a large number of imaged tumors were given to a board of experienced physicians and asked to rate their heterogeneity. Shown in Figure 4, tumors with irregular shape and non-uniform coloring were rated more highly heterogeneous by the physicians and also received higher heterogeneity factors from their quantitative assessment.[5]

This method was replicated using Python and Pydicom, a package created to allow data to be extracted DICOM files, the standard format for medical images [See Appendix I for code]. Following the completion of the code, we tested this metric on known homogeneous and known heterogeneous subjects, via both phantom and simulated phantom studies, for validity and viability of use in a clinical setting. Concerning phantom images, we chiefly utilized water-filled torso phantoms, pictured in Figure 2, and isolated the small reservoirs of high-concentration FDG as our regions of interest. The reservoirs contained four times the concentration of FDG relative to the background, and images were taken using TORSO PET routine on PET-MRI machines and 100-150lbs routine on PET-CT. All simulated data was created in Python.
Figure 4: examples of tumors with measurements of quantified heterogeneity. Heterogeneity factors can be found on the upper-right corner of each box, and image size can be found in each bottom-left corner.\[5\]
VII. Results

Heterogeneity measurements on phantom test objects produced heterogeneity values several orders of magnitude larger than measured in qualitatively similar objects by Brooks. The Brooks Study analyzed images stored using only 48 shades of gray. However, most imaging equipment outputs data in the form of 12 or 16-bit images, corresponding to 4096 and 65536 allowed shades of gray respectively. This allowed for much larger inter-pixel intensity differences to occur, which resulted in large heterogeneity factors. This large magnitude difference was resolved with a simple normalization, which can be generalized to images of any bit depth.

\[ \zeta' = \frac{\zeta}{2^{\text{Bit Depth}}} \]

Figure 5: heterogeneity measurements from a PET-CT phantom, showing normalized and unnormalized result
The method of Brooks was also cited as size-invariant, but our results show consistent size dependence, regardless of imaging modality or presence of image attenuation correction. Illustrated below, the PET images of the same phantom were taken on both PET-MRI and PET-CT machines, with and without attenuation correction.

**Figure 6: labeled torso phantom, PET-CT with AC**
Figure 7: labeled torso phantom, PET-CT without AC

Figure 8: labeled torso phantom, PET-MRI with AC
Figure 9: labeled torso phantom, PET-MRI without AC

Figure 10: data from figures 6-9, sorted by reservoir size
To further illustrate this phenomenon, an array of simulated phantoms of varying size were created and superimposed with Gaussian noise of varying standard deviation. 100 phantoms were tested at each size and noise intensity, and results again showed the same increasing heterogeneity factor.

Despite the unexpected emergence of this size dependence, results for images of visibly higher heterogeneity and pixel variation were always rated more heterogeneous than others with higher apparent uniformity.

Figure 11: simulated phantom data, varying size and noise amplitude
Figure 12: simulated heterogeneous phantom, noiseless

Figure 13: simulated heterogeneous phantom, with noise
VIII. Discussion

We believe that the emergence of this size dependence is a direct consequence of the differing image noise amplitude and pixel size. Standard deviation of pixel intensity for each phantom reservoir was typically on the order of 5% of the maximum intensity. With noise of this magnitude, along with the low resolution typical to PET images, smaller ROI can become obscured and appear to have higher uniformity.

Despite issues with characterization of size dependence, this method of heterogeneity quantification matches very well with visual assessment, by all counts, and results have remained fairly consistent through all of the computational tests incurred thus far. However, in all images of the torso phantom, the smallest reservoirs were typically unable to be tested due to their low relative intensity compared to background signal. Further investigation into minimum ROI size could reveal more about the size constraints with this method, as well as any technical issues with attenuation correction.

Having seen validity in this method’s ability to quantify heterogeneity, the next step would be to focus on patient data and attempt to find some sort of prognostic significance. Many other groups that have attempted to quantify heterogeneity have also expressed a desire to use this type of measurement to develop a prediction of treatment outcome. In future work, we intend to use this metric to analyze patient data, where images throughout the entire duration of the treatment have been taken. With this information, we hope to draw out meaningful information relating to prognostic significance from the heterogeneity
measurement. We anticipate imaged tumors with high heterogeneity to be correlated with poorer prognoses, but this could emerge in a wide variety of forms, such as bad response to treatment, high rates of recurrence, higher rates of mortality, etc.
IX. Conclusion

Tumor heterogeneity has been a hot topic among physicians for years due its high degree in unpredictability and effect on treatment response. This method of quantifying heterogeneity has held validity through all of our imaging assessments using PET-CT, PET-MRI, and simulated data. However, to make this heterogeneity statistic more quantitative, we found that a normalization to bit depth is necessary and, although described as size-invariant, real-world noise amplitude, pixel size, and resolution can create a size dependence in small objects. Because this method has been verified to correlate strongly with a physician’s qualitative assessment of heterogeneity, it is reasonable to expect strong correlation with clinical outcomes. We anticipate that this statistic will be useful to characterize and develop meaningful prognostic predictions of heterogeneous tumors imaged in human patients.
X. Acknowledgements

I would like to thank Dr. David Jordan, Dr. Norbert Avril, and the Department of Radiology at University Hospitals for their help in planning, data collection, and data analysis. I would also like to thank Dr. Frank Brooks for his advice in recreating his methods.
XI. Citations


XII. Appendix I: Code

```python
import scipy.integrate as integ
import dicom

def heterogeneity(data, low):
    thresh = minint(data, low)  # set minimum allowed pixel threshold
    delta_i = []
    for x in range(0, shape(data)[0]):
        for y in range(0, shape(data)[1]):
            if (data[x][y] > 0):
                for j in range(x+1, shape(data)[0]):
                    if (data[j][y] > 0):
                        bres = get_line(x, y, i, j)  # create bresenham line between pixel pair
                        rel = relint(data, bres, thresh)  # finds relative intensity and length for bres line
                        delta_i = append(delta_i, rel)  # relative intensity saved here
    stuff, bins = histogram(delta_i.T[1], bins = round(max(delta_i.T[1])))  # creates sorting increment based on bres line length
    av = avgrelint(delta_i, bins)  # finds average relative intensity for each group of lengths
    z = integ.trapz(av, bins/bins[-1])
    return z, z/4096

def dis(p1, p2):  # basic distance equation
    r = sqrt((p1[0]-p2[0])**2 + (p1[1]-p2[1])**2)
    return r

def minint(pixarray, low):  # finds minimum allowed pixel intensity
    w = reshape(pixarray, (shape(pixarray)[0]*shape(pixarray)[1],1))
    m = max(w)
    return low*m

def relint(inten, bres, thresh):
    m = inten[bres[0][0], bres[0][1]]  # intensity of first point in bres line
    n = inten[bres[-1][0], bres[-1][1]]  # intensity of last point in bres line
    if (m < thresh):
        m = 0
    if (n < thresh):
        n = 0
delta_i = 0
    L = dis((bres[0][0], bres[0][1]), (bres[-1][0], bres[-1][1]))  # total length of bres line
    for i in range(1, len(bres)):
        l = inten[bres[i][0],bres[i][1]]  # intensity of intermediate point on line
        if (l < thresh):
            l = 0
        delta_i += abs(m - l + (n - m)*dis((bres[0][0], bres[0][1]), (bres[i][0], bres[i][1])))/L  # combination of eqns 1 and 2 in paper
    delta_i = delta_i / L  # normalized to length of bres line
    return delta_i, L

def avgrelint(delta_i, bins):  # groups relative intensities by length of bres line, averages
    usedind = []
```

aver = zeros((len(bins)))
for i in range(0, len(bins)):
    ind = where(delta_i.T[1] - bins[i] <= 0)[0]
    n = 0.0001
    for j in range(0, len(ind)):
        if ind[j] not in usedind:
            aver[i] += delta_i.T[0][ind[j]]
            n+=1
    aver[i] = aver[i]/n
    usedind = append(usedind, ind)
return aver

def get_line(x1, y1, x2, y2):
    # finds bresenham line
    points = []
    issteep = abs(y2-y1) > abs(x2-x1)
    if issteep:
        x1, y1 = y1, x1
        x2, y2 = y2, x2
        rev = False
    if x1 > x2:
        x1, x2 = x2, x1
        y1, y2 = y2, y1
        rev = True
    deltax = x2 - x1
    deltay = abs(y2-y1)
    error = int(deltax / 2)
    y = y1
    ystep = None
    if y1 < y2:
        ystep = 1
    else:
        ystep = -1
    for x in range(x1, x2 + 1):
        if issteep:
            points.append((y, x))
        else:
            points.append((x, y))
        error -= deltay
        if error < 0:
            y += ystep
            error += deltax
    # Reverse the list if the coordinates were reversed
    if rev:
        points.reverse()
    return points