Applications and limitations of radiomics

To cite this article: Stephen S F Yip and Hugo J W L Aerts 2016 Phys. Med. Biol. 61 R150

View the article online for updates and enhancements.

Related content
- Use of registration-based contour propagation in texture analysis for esophageal cancer pathologic response prediction
  Stephen S F Yip, Thibaud P Coroller, Nina N Sanford et al.
- Stability of radiomic features in CT perfusion maps
  M Bogowicz, O Riesterer, R A Bundschuh et al.
- Significance of the impact of motion compensation on the variability of PET image features
  M Carles, T Bach, I Torres-Espallardo et al.

Recent citations
- Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT
  Wenbing Lv et al
- Magnetic Resonance Spectroscopy, Positron Emission Tomography and Radiogenomics—Relevance to Glioma
  Gloria C. Chiang et al
- Yong Du and Habib Zaidi

This content was downloaded from IP address 54.70.40.11 on 30/03/2018 at 05:46
Topical Review

Applications and limitations of radiomics

Stephen S F Yip\(^1\) and Hugo J W L Aerts\(^{1,2}\)

\(^1\) Department of Radiation Oncology, Brigham and Women’s Hospital, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
\(^2\) Department of Radiology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA

E-mail: Hugo_Aerts@dfci.harvard.edu

Received 21 July 2015, revised 29 February 2016
Accepted for publication 5 April 2016
Published 7 June 2016

Abstract

Radiomics is an emerging field in quantitative imaging that uses advanced imaging features to objectively and quantitatively describe tumour phenotypes. Radiomic features have recently drawn considerable interest due to its potential predictive power for treatment outcomes and cancer genetics, which may have important applications in personalized medicine. In this technical review, we describe applications and challenges of the radiomic field. We will review radiomic application areas and technical issues, as well as proper practices for the designs of radiomic studies.

Keywords: PET, quantitative imaging, radiomics

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

1. Introduction

Non-invasive medical imaging, such as magnetic resonance (MR) imaging, computed tomography (CT), and positron emission tomography (PET), is routinely used for assessing tumour and anatomical tissue characteristics for cancer management (Buckler et al 2011, Kurland et al 2012). Furthermore, imaging can potentially provide valuable information for personalized medicine that aims to tailor treatment strategy based on the characteristics of individual patients and their tumours.

Molecular characterization using genomics, proteomics, and metabolomics information has been the main focus of personalized therapy. However, spatial and temporal intratumoural heterogeneity that arises from regional variations in metabolism, vasculature, oxygenation, and gene expression is a common feature of malignant tumours (Maley et al 2006, Marusyk et al 2012, Chicklore et al 2013, Fisher et al 2013). Random samples of tumour tissues acquired through invasive biopsy for molecular characterization may thus fail to accurately represent...
the landscape of the biological variation within tumours (Gerlinger et al 2012). On the other hand, the entire tumour can be sampled non-invasively and repeatedly with medical imaging.

In particular, studies have hypothesized that tumour characteristics at the cellular and genetic levels are reflected in the phenotypic patterns that can captured with medical images (Henriksson et al 2007, Diehn et al 2008, Basu et al 2011, Yang and Knopp 2011). Several studies have shown these associations across imaging modalities. For example, in MR, growing brain tumours that cause a shift in midline structures due to normal tissue compression, known as the mass effect, are found to be strongly correlated with proliferation gene-expression (Diehn et al 2008). Yamamoto et al (2014) found that lung tumours with anaplastic lymphoma kinase (ALK) mutations appeared to have larger pleural effusion and no pleural tails on CT images (Yamamoto et al 2014). Contrast-enhanced CT images revealed that the mutation status of von Hippel–Lindau (VHL) in renal cell carcinoma is significantly correlated with the ‘gross appearance of intratumoural vascularity’, ‘well-defined tumour boundaries’, and ‘enhancement of nodular tumour’ (Karlo et al 2014). In PET imaging, $^{18}$F-fluorodeoxyglucose ($^{18}$F-FDG) uptake is related to the number of viable cancer cells, tumour histopathology, and a number of biological processes that support the continuous growth of the tumour (Higashi et al 1993, Haberkorn et al 1994, Rajendran et al 2004, Fanchon et al 2015). Studies have therefore proposed that tumour heterogeneity may be associated with the non-uniform distribution of $^{18}$F-FDG (Henriksson et al 2007, Tixier et al 2011, Chicklore et al 2013).

Despite the promise of medical imaging to assess tumour heterogeneity (or genetic), imaging features are often assessed visually and described qualitatively by radiologists or nuclear medicine physicians. Subjective descriptions of tumour imaging phenotypes (e.g. ‘large necrotic core’, ‘highly speculated’, and ‘moderate heterogeneity’). However, these visual assessments can suffer from a large intra and inter-observer variability (de Jong et al 1995, Mussurakis et al 1996, Wetzel et al 2002, Davnall et al 2012, Tixier et al 2014b). Therefore, it is important to objectively and reproducibly quantify various imaging features that may reveal the underlying biology of tumours.

Radiomics uses the high-throughput extraction of advanced quantitative features to objectively and quantitatively describe tumour phenotypes (figure 1). These features, termed radiomic features, are extracted from medical images using advanced mathematical algorithms to uncover tumour characteristics that may fail to be appreciated by the naked eye (Lambin et al 2012, Chicklore et al 2013, Aerts et al 2014, Cook et al 2014, Buvat et al 2015, Rahmim et al 2016). Radiomics may thus provide great potential to capture important phenotypic information, such as intra-tumour heterogeneity, subsequently providing valuable information for personalized therapy. In this review, we will review radiomic applications and technical limitations, as well as proper practices for the designs of radiomic studies.

2. Potential applications of radiomics

Numerous radiomic features, such as size and shape based–features, descriptors of the image intensity histogram, descriptors of the relationships between image voxels (e.g. gray-level co-occurrence matrix (GLCM), run length matrix (RLM), size zone matrix (SZM), and neighborhood gray tone difference matrix (NGTDM) derived textures), textures extracted from wavelet and Laplacian of Gaussian (LoG) filtered images, and fractal features, can be extracted from the medical images (Haralick et al 1973, Galloway 1975, Pentland 1984, Amadasun and King 1989, Davnall et al 2012, Thibault et al 2013, Aerts et al 2014, Rahmim et al 2016).

Radiomic features not only provide an objective and quantitative way to assess tumour phenotype, they have also found a wide-range of potential applications in oncology. For example,
Radiomic features have shown promise in the prediction of treatment response, differentiating benign and malignant tumours, and assessing cancer genetics in many cancer types. We will review the potential application of the radiomic features.

2.1. Prediction of treatment response and outcomes

MR studies have shown that intensity histogram-based radiomic features are potentially useful for predicting cancer response to treatment (Johansen et al. 2009, Baek et al. 2012, Shukla-Dave et al. 2012, King et al. 2013, Peng et al. 2013). In pre-clinical model, (Foroutan et al. 2013) observed that mice with sarcomas treated with combinations of MK1775, a cell cycle checkpoint inhibitor, and gemcitabine showed a substantial change in the (apparent diffusion coefficient) ADC histogram skewness, kurtosis, entropy, and average ADC shortly after treatment compared to the untreated control group. In human patients with head-and-neck cancer, tumours that responded poorly to chemoradiotherapy demonstrated a significantly greater increase in average ADC and higher values in kurtosis and skewness on mid-treatment diffusion weighted MR (DW-MR) than tumours with a better therapeutic response (King et al. 2013). K-trans is a measure derived from dynamic CT MR images and measures the diffusion of an intravascular contrast agent into the extracellular space. The skewness of K-trans was found to be a promising predictor of progression free survival and overall survival of patients with stage IV head-and-neck cancer (Shukla-Dave et al. 2012). The findings of these aforementioned studies may support the notion that therapy induced changes in tumour

Figure 1. Radiomic workflow. (Top) Various radiomic features, such as shape/size-based, histogram-based, filtered-based, and textural features, can be extracted from the medical images within the tumours. (Bottom) The radiomic features are then compared with the clinical and genomics data.
microenvironment and composition can be potentially described by changes in the intensity-histogram shape.

In PET imaging, standardized uptake value (SUV) measures, such as the maximum SUV (SUV\text{\text{max}}) and mean SUV obtained within a tumour (SUV\text{\text{mean}}), are commonly used for tumour characteristic quantification. High baseline SUV uptake is often thought to be associated with aggressive tumour behavior and poor prognosis (Rizk \textit{et al} 2006, Zhang \textit{et al} 2011, Higgins \textit{et al} 2012). However, as previously mentioned, SUV\text{\text{max}} and SUV\text{\text{mean}} are inadequate for describing the heterogeneous distribution of [\text{18F}]FDG uptake (van Velden \textit{et al} 2011, Marusyk \textit{et al} 2012, Cheng \textit{et al} 2013b).

Recently, radiomic textural features have drawn considerable interest due to their potential to describe distinctive tumour phenotype (‘appearance’) that may be driven by underlying genetic and biological variability. In particular, they were demonstrated to outperform simple SUV measures, such as SUV\text{\text{max}} and SUV\text{\text{mean}}, in treatment outcome prediction (Eary \textit{et al} 2008, El Naqa \textit{et al} 2009, Tixier \textit{et al} 2011, Yang \textit{et al} 2013). For example, Cook \textit{et al} (2013) compared the predictive power of maximum and mean SUV and four NGTDM derived textures in fifty-three non-small cell lung cancer (NSCLC) patients (Cook \textit{et al} 2013). They found that NGTDM derived coarseness, busyness, and contrast could better differentiate between responders and nonresponders to chemoradiotherapy than the aforementioned SUV measures. Furthermore, coarseness was found to be an independent predictor of patient overall survival. Zhang \textit{et al} (2014) built several multivariate models to predict pathologic response to preoperative chemoradiotherapy in twenty esophageal cancer patients. They found that models constructed with combined radiomic features significantly improved the pathologic response prediction compared to models built with maximum SUV, metabolically active tumour volume and longest diameter.

For CT imaging, Aerts \textit{et al} (2014) assessed the prognostic values of 440 shape- and intensity-based and textural features. They identified features that were predictive of patients’ survival on a discovery dataset consisting of >420 lung cancer patients. The prognostic value of features were then validated on three independent datasets, including one lung cancer (225 patients) and two head-and-neck cancer (231 patients) cohorts. Their results not only confirmed the potential use of radiomic features in outcome prediction and describing intratumoural heterogeneity, but also showed that prognostic ability may be transferred from one disease type to another (i.e. from lung to head-and-neck cancer). On the other hand, Parmar \textit{et al} (2015b) observed that not all radiomic features that significantly predicted lung cancer patients’ survival also predicted survival in head-and-neck cancer patients and vice versa (Parmar \textit{et al} 2015b). Their results thus suggested that some radiomic features may be cancer-specific.

Studies have found a strong association between contrast-enhanced CT (CE-CT) and heterogeneity of the tumour vasculature (Tateishi \textit{et al} 2002, Kim \textit{et al} 2005). Tixier \textit{et al} (2014a) observed that tumour blood flow measured on CE-CT was significantly correlated with the metabolically active tumour volume. CE-CT radiomic therefore provides great potential for quantifying complex tumour phenotype arising from angiogenesis in cancer. For example, Hayano \textit{et al} (2014) hypothesized that if the fractal dimension extracted from CE-CT is useful in describing tumour heterogeneity, then the measure may also be useful for predicting patient survival in hepatocellular carcinoma (Hayano \textit{et al} 2014). They found that the patients with longer survival often had lower fractal dimensional on the arterial phase CE-CT image.

Furthermore, radiomic features can be potentially applied to assess the metastatic potential of tumours. Coroller \textit{et al} (2015) identified thirty-five CT radiomic features to be significant predictors of distant metastasis and six features to be predictors of survival in 182 lung cancer
patients (Coroller et al 2015). They concluded that the radiomic features they identified may be useful for early indication of cancer patients that will have a high risk of developing distant metastasis, thus allowing physicians to better adapt treatment plans for individual patients. Recently, Vallieres et al (2015) showed that the combination of MR and $^{18}$F-FDG-PET textural features can better predict the risk of lung metastases in soft-tissue sarcomas than the features acquired from a single modality (Vallieres et al 2015).

2.2. Tumour staging

Many radiomic features were shown to be able to significantly differentiate between early and advanced stage diseases. For example, in a PET radiomic study by (Dong et al 2013), forty esophageal cancer patients were staged according to the American Joint Committee on Cancer (7th edition). SUV$_{\text{max}}$, GLCM-entropy, and GLCM-energy were found to be significantly correlated with T and N stage. In particular, a GLCM-entropy value $>4.70$ could accurately identify tumours with stages above stage IIb (Dong et al 2013). In a recent study, Mu et al (2015) classified forty-two cervical cancer patients into early stage (stage I and II) and advanced stage (stage III and IV) using PET-based radiomic features (Mu et al 2015). RLM-run percentage texture was found to be most associated with cervical tumour stage. Moreover, CT-based fine textures derived from LoG filtered CT images were found to predict lung tumour stages above stage II (Ganeshan et al 2010). Early identification of tumour stage using radiomic features may assist physician to better stratify patients, subsequently selecting the best treatment for individual patients.

2.3. Tissue identification

Radiomic features have also been shown to be useful in discriminating between malignant and other tissues in many disease types. In the 1990s, GLCM textures derived from a 2D slice of T1- and T2-weighted MR images were first reported to be potentially useful for tissue differentiation, with the ability to differentiate brain tumour tissue, edema, cerebrospinal fluid (CSF), white matter, and gray matter, in patients with brain cancer (Lerski et al 1993, Kjær et al 1995). Mahmoud-Ghoneim et al (2003) demonstrated that GLCM textures computed within a 3D volume of the MR images outperformed 2D textures in separating necrosis and edema from solid tumours (Mahmoud-Ghoneim et al 2003). Besides brain tumours, Nie et al (2008) showed that combining shape-based, volume-based, and GLCM textural features of MR using an artificial neural network (ANN) may be used to differentiate malignant from benign tumours in breast cancer (AUC $\geq 0.80$) (Nie et al 2008). Furthermore, they also observed that benign tumours had smoother boundaries, rounder shape, and lower image intensity than malignant tumours.

CT-based radiomic features have been used to classify a pulmonary nodule as benign or malignant (McNitt-Gray et al 1999, Kido et al 2002, Petkovska et al 2006, Way et al 2006). Pulmonary nodules could be due to other diseases (e.g. tuberculosis and fungal infection) than cancer. Kido et al (2002) showed that the fractal dimensions for bronchogenic carcinomas were significantly smaller than pneumonias and tuberculosis ($p < 0.0001$). Petkovska et al (2006) showed that GLCM textures extracted from CE-CT can accurately identify malignant from benign nodules, while visual inspection by three experienced radiologists performed worse in malignant-benign nodule differentiation. Combining shape-, size, and histogram-based features has been show to improve the differentiation between malignant and benign nodules (Way et al 2006).
Furthermore, a study by (Xu et al 2014) developed a computer aided diagnosis method with combined CT- and PET-based textures for differentiating malignant and benign lesion in various tumour sites. $^{18}$F]FDG uptake of malignant lesions were observed to be more heterogeneous than the benign tissues. Compared with the histological diagnosis of the lesions, the classification results of their texture-based diagnosis method achieved sensitivity, specificity, and accuracy $>75\%$ (Xu et al 2014). Yu et al (2009) assessed the ability of 14 PET and 13 CT-based textures in delineating primary and nodal tumours from normal tissues. The sensitivity, specificity, and accuracy of the delineation results based on the radiomic textural features were $>95\%$ comparing to the tumour contours manually segmented by three radiation oncologists (Yu et al 2009).

2.4. Assessment of cancer genetics

Many studies have shown that there is a strong relationship between imaging features and the underlying tumour genetics, which may provide a biological basis for the clinical applications of radiomics. MR-based volumetric features are often observed to be associated with somatic mutations and genetic expression of brain tumours (Diehn et al 2008, Ellingson et al 2013, Naeini et al 2013, Gutman et al 2015). For instance, Ellingson et al (2013) observed that MGMT unmethylated glioblastoma (GBM) usually had smaller volumes of T1-contrast enhanced and T2-FLAIR hyperintensity than methylated GBM. In a recent study by Gutman et al (2015), volumetric measures, such as contrast enhancing volume, necrosis volume, and total tumour volume, were found to significantly predict GBM mutations, including TP53, NF1, EGFR, RB1, and Platelet-derived growth factor receptor alpha (PDGFRA).

In CT imaging, Aerts et al (2014) found that radiomic features related to shape and wavelet features describing the heterogeneous phenotype of lung tumours were found to be significantly associated with cell cycle pathway, suggesting that highly proliferative tumours demonstrate complex imaging patterns. Moreover, various biological mechanisms may be described by different radiomic features as the features were found to be related with different biological gene sets, including DNA recombination and regulation of DNA metabolic processes (Aerts et al 2014).

Nair et al (2012, 2014) investigated the association between PET-SUV histogram radiomic features and various NSCLC genes and gene expressions in cohorts consisting of $>300$ patients (Nair et al 2012, 2014). Features such as skewness, $\text{SUV}_{\text{max}}$, $\text{SUV}_{\text{mean}}$, median of the SUV histogram were strongly correlated with several gene signatures and expressions (e.g. NF-$\kappa$B) that are related to patient survival (Nair et al 2012, 2014).

As numerous radiomic features can be extracted from medical images, the studies mentioned in this section play an important role in identifying only a subset of features that might be most relevant to the underlying tumor biology and genetics. However, how the tumor pathophysiological processes give rise to imaging phenotypes that can be quantified by radiomic features remain unclear. Future studies would need to investigate these associations to further elucidate the biological meaning of the radiomic features.

3. Factors that affect radiomic features quantification

3.1. Acquisition modes, reconstruction parameters, smoothing, and segmentation thresholds

Despite the wide range of potential applications, radiomic feature quantification may be sensitive to a number of technical factors. For example, Galavis et al (2010) assessed the variability of 50 PET radiomic features due to different acquisition modes, matrix sizes, post-filtering
widths, reconstruction algorithms and iteration numbers (Galavis et al 2010). Of these features, forty were shown to have substantial variability with a relative difference of >30%. Only four features, including intensity-histogram derived entropy and energy, GLCM-maximal correlation coefficient, RLM-low gray level run emphasis, were found to have variability <5%. The textures that are sensitive to acquisition modes and reconstruction parameters are thus not recommended for radiomic applications, such as malignant and benign tissue differentiation (Galavis et al 2010). Yan et al (2015) identified features, including histogram-based entropy, normalized GLCM-inverse difference moment and inverse difference, RLM-low gray run emphasis and high gray run emphasis, and SZM-low gray zone emphasis, were robust to different PET image reconstruction settings (Yan et al 2015). These features may thus be useful for radiomic studies. However, (Galavis et al 2010, Yan et al 2015) did not investigate or elaborate on why certain radiomic features were more sensitive to the others. This may need to be further investigated.

On the other hand, (Doumou et al 2015) investigated the effect of PET image post-filtering width (noise smoothing) on feature quantification. They found that the radiomic features were generally insensitive to variations in filter width.

Accurate delineation of tumour volumes is crucial for the computation of radiomic features. Manual delineation of tumour volume is not only time-consuming, but can also be affected by inter-observer variability. Radiomic studies often recommend using automatic and semi-automatic methods for tumour volume delineation over manual contouring (Hatt et al 2009, Velazquez et al 2013, Parmar et al 2014, Yip et al 2016). For example, Velazquez et al (2013) compared the accuracy of manual and semiautomatic region growing tumour contouring methods on CT images. They found that the semiautomatic contouring method was better associated with gold-standard tumour sizes that were measured by surgical resection. Moreover, Parmar et al (2014) found that CT-based radiomic features were more stable when computed from a semiautomatic contouring method than from manual contours (Parmar et al 2014). A recent study investigated metabolic tumour volume auto-segmentation thresholds (45–60% of the maximum SUV) on the precision of PET-based radiomic texture quantification (Doumou et al 2015). The authors concluded that the variation in image segmentation thresholds only have small effects on the quantification, suggesting metabolic tumour volume may be precisely defined by thresholding.

Hatt et al (2011) and Cheebsumon et al (2012) found that lung tumour size computed with PET-based tumor delineation methods, such as fixed and adaptive thresholds, are in better agreement with surgical resection while manual contouring on CT images significantly overestimated the pathological tumour size. However, more advanced delineation algorithms, such as fuzzy locally adaptive Bayesian (FLAB), are recommended for larger lung tumors as simple threshold-based methods may result in underestimation of the metabolically active tumor region (Hatt et al 2011).

3.2. Reproducibility of radiomic features

While tumour heterogeneity can be potentially quantified using numerous radiomic features extracted from medical images, many features are often found to be unstable between imaging scans acquired within weeks–even minutes of each other (Tixier et al 2012, Leijenaar et al 2013, Balagurunathan et al 2014a, 2014b, van Velden et al 2014). Balagurunathan et al (2014a, 2014b) assessed the intra-class correlation coefficient (ICC) of 219 radiomic features extracted from a test and retest CT scan in lung cancer patients which were acquired 15 min apart. Of these features, only 66 of them were found to have ICC ≥ 0.90 across the test and retest scans, suggesting that a large number of features may be unreliable. Tixier
et al (2012) studied the reproducibility of 25 radiomic features between repeated PET scans acquired within 4 d of each other. GLCM-entropy and homogeneity, SZM-zone and intensity variability were not only found to be significant predictors of treatment response (Tixier et al 2011), but also exhibited the highest reproducibility (Tixier et al 2012). Leijenaar et al (2013) studied the stability of nearly 100 radiomic features to repeated PET images (1 d apart) and inter-observer variability in tumour delineation in lung cancer patients. They found that the PET-based features that are stable between repeated scans were also more robust to inter-observer variability, suggesting that features with poor reproducibility may also be sensitive to other factors and are thus not recommended.

However, to our knowledge, the repeatability of MR-based radiomic features has not been investigated. Understanding the stability of MR-based radiomic features between test and re-test scans can help identifying reliable features for radiomic applications, and thus would be a valuable future study.

3.3. Image discretization (resampling) schemes

Prior to radiomic feature computation, voxel intensities within tumour volumes need to be discretized to a limited range of intensity values in order to efficiently and practically compute the radiomic features (Cheng et al 2013b, Leijenaar et al 2013). For instance, an image with 1024 discrete intensity values will yield a $2^{1024} \times 2^{1024} \times 2^{1024}$ GLCM (Haralick et al 1973), which can be computationally intensive. The range of intensity values of an image thus needs to be reduced and limited for efficient radiomic feature computation. In PET radiomic studies, the most commonly used discretization scheme is to first normalize the medical image by the relative difference of the maximum and minimum intensity values within tumours, and then resample the voxel intensities to $2^N$ values (i.e. $2^N$ number of bins), where $N$ ranges from 3 to 8 in literature (Orlhac et al 2014, Tiévi et al 2014a, Doumou et al 2015, Mu et al 2015). Studies have shown that both the quantity and prognostic value of radiomic features, particularly GLCM-entropy, SZM-size zone high gray emphasis and SZM-size zone non-uniformity derived from PET images, can be substantially influenced by the number of discrete values ($2^N$) (Cheng et al 2013a, Orlhac et al 2014, Doumou et al 2015). At least $2^5 = 32$ discrete values is recommended to properly quantify tumour heterogeneity with PET-based radiomic features (Orlhac et al 2014). However, the Spearman correlation coefficient of tumour volume and GLCM-entropy was observed to be $>0.85$ for resampling values over 64, suggesting that textural features computed with resampling values over 64 may not provide additional prognostic information compared with the tumour volume (Hatt et al 2015). Therefore, Hatt et al (2015) limited the number of discrete bins to 64 for PET-based texture computation.

Alternatively, the voxel intensity range can be discretized into equally spaced bins with a fixed bin width (Leijenaar et al 2013, 2015b). For instance, Leijenaar et al (2015b) compared two resampling strategies (i.e. fixed number of bins and fixed bin width) and found that quantification of radiomic features were more robust to a change in bin size than to a change in the number of bins (Leijenaar et al 2015b). They concluded that the resampled PET image voxel intensity (or SUV) using a fixed bin width may be more appropriate for clinical case studies. This is because resampled PET voxel intensity using the alternative (i.e. a fixed number of bins) implicitly assumes that the tumour images of all patients have the same SUV range, which is usually not the case.

Another resampling strategy is to determine the bin size for each tumour image according to the Freedman–Diaconis rule (bin size $= 2 \times \text{IQR} / N^{1/3}$), where IQR is the interquartile SUV range and $N$ is total number of voxels that the tumours are composed of (Brooks and Grigsby
Comparison of different resampling strategies may be important to understand their effect on radiomic features in treatment outcome prediction.

3.4. Computation of radiomic features

The computations of radiomic features, even with the same features names, may be implemented differently in radiomic studies. For example, GLCM can be calculated either by averaging the values of the matrices computed for 13 distinct directions or a single matrix that accounts for tumour co-occurrence information in all 13 directions (Hatt et al 2015). Textural features can be extracted from the largest cross-sectional (axial) slice of the tumour boundary (2D textures) or extracted from the entire tumour volume (3D textures) (Ng et al 2013, Fave et al 2015). The impact of different feature implementation/computation methods on the predictive values of radiomic features needs to be carefully studied.

3.5. Respiratory motion

The accurate quantification of radiomic features can be hindered by respiratory motion in lung cancer patients (Yip et al 2014). Lung motion can lead to a reduction in the measured activity in tumor and other tissues due to insufficient data acquisition and limited reconstruction techniques in static PET images (3D PET) (Nehmeh et al 2002, Aristophanous et al 2012, Huang and Wang 2013). 4D PET imaging gates PET image acquisition with respiratory motion to improve PET image quality (Nehmeh et al 2002, García Vicente et al 2010, Didierlaurent et al 2012). Yip et al (2014) investigated the influence of the lung tumour motion on radiomic textures. They observed that the radiomic textural features, blurred out by respiratory motion during 3D-PET acquisition, can be better resolved by 4D-PET imaging. 4D-PET textures may have better prognostic value as they are less susceptible to tumour motion although the hypothesis needs to be investigated in the future.

3.6. Tumor size and intratumoural heterogeneity

Intratumoural heterogeneity for small tumour volumes may not be accurately quantified due to the partial volume effect resulting from limited PET resolution (Soret et al 2007, Hatt et al 2013). Therefore, many studies often exclude tumours with volumes <3–5 cm$^3$ from radiomic analysis (Orlhac et al 2014, Hatt et al 2015). To estimate the minimum tumour volume needed for texture computation, Brooks and Grigsby (2014) extracted GLCM-entropy from PET images in 70 cervical cancer tumours (Brooks and Grigsby 2014). Using probability theory, they found that GLCM-entropy computed for tumours <45 cm$^3$ were strongly correlated with tumour size, and therefore may not accurately measure intratumoural heterogeneity. However, their conclusion was based on theoretical analysis, one radiomic texture, and a single tumour type.

Hatt et al (2015) computed four prognostic radiomic textures features on 555 PET images acquired from multiple cancer centers consisting of breast, cervical, NSCLC, esophageal, and head-and-neck tumours (Hatt et al 2015). The added prognostic value of the textures and their correlation to tumour volume were investigated. Both radiomic textures and tumour volume were observed to be independent predictors of survival for patients with bigger tumours, whereas the added value of textures in predicting survival was minimal for small tumours. They observed a strong correlation between textural features and tumour size for tumours with volumes less than 10 cm$^3$. The results of Hatt et al (2015)’s study suggest that radiomic textures have no added value in outcome prediction for tumours <10 cm$^3$. However, instead
of excluding tumours with volume < 10 cm³ in the future radiomic studies, they recommended that the correlation of the radiomic features and tumour volume should be always reported to highlight if the features provide independent or redundant information (Hatt et al. 2015).

### 3.7 The silver lining and the need for standardization

Besides the aforementioned factors, there are other factors, such as metal artifacts in CT images (Leijenaar et al. 2015a), CT x-ray tube peak voltage and current (Fave et al. 2015), that may also affect radiomic feature quantification. As CT images are often employed for attenuation correction of PET and single-photon emission computed tomography (SPECT) images, factors that affect the quality of the CT images can also impact the quantification of features extracted from the PET and SPECT images. Despite the potential impact of these factors on quantification, strong prognostic signals of the features could still be found (Cheng et al. 2013a, 2014, Cook et al. 2013, Aerts et al. 2014, Coroller et al. 2015, Leijenaar et al. 2015a, Parmar et al. 2015b). While harmonization and standardization for imaging acquisition and feature computation may lead to more consistent findings in radiomic studies across institutions, the technical factors that affect the radiomic feature quantification may not be reduced (Boellaard 2011, Nyfflot et al. 2015). For example, in harmonization, as some PET systems fail to fully resolve small objects due to limited resolution (partial volume effect), additional smoothing steps are thus required for images acquired by certain PET systems, even with high resolution and sensitivity (Boellaard et al. 2015). Thus, the impact of harmonization and standardization on the quantification and predictive values of radiomic features would be an important topic of future investigations for the field of radiomics. Of equal importance, standardization for proper statistical practice and study designs for current radiomic studies also need to be considered.

### 4. False positive discovery rate and proper study design

Many studies examined the prognostic value of radiomic features based on retrospective analysis of small patient datasets (< 50 patients) (Tixier et al. 2011, Dong et al. 2013, Tan et al. 2013, Bundschuh et al. 2014, Zhang et al. 2014). These retrospective studies are important for providing rationale (or proof-of-concept) for further investigation of radiomic features as imaging biomarkers and surrogates for intratumoural heterogeneity. However, it is not uncommon that the number of examined radiomic features is much greater than the number of patients, which can lead to feature selection bias and false positive results (Alic et al. 2014, Chalkidou et al. 2015). To demonstrate this bias, Chalkidou et al. (2015) randomly generated 100 features and assessed the association between the features and survival data extracted from a study by Ganeshan et al. (2012) consisted of only 21 esophageal cancer patients (Ganeshan et al. 2012). Ten random features were found to accurately identify patients surviving a follow-up period of over 2 years with the area under the receiver-operating-characteristics curves (AUC) of 0.68–0.80.

Ideally, an external validation dataset is required to confirm the prognostic value of the radiomic features to avoid optimism based on false positive results (Steyerberg et al. 2010, Lambin et al. 2013, Aerts et al. 2014, Chalkidou et al. 2015). However, acquiring a validation dataset is not always feasible due to high cost, requirement of excessive effort, differences in data collection practice and privacy issues between institutes (Lambin et al. 2013).

As a rule of thumb, to reduce the false discovery rate, 10–15 patients are needed for each examined radiomic feature (Chalkidou et al. 2015). As many of the radiomic features are
highly correlated, radiomic studies should avoid including strongly correlated features that may provide redundant information about tumour characteristics (Orlhac et al 2014, Mu et al 2015). For analyses where large numbers of radiomic features are studied, the significant values (p-values) should be corrected for multiple hypotheses testing using the Holm–Bonferroni method or a false-discovery rate (FDR) controlling procedure, such as the Benjamini–Hochberg method (Alic et al 2014, Chalkidou et al 2015). For example, the Benjamini–Hochberg procedure has been used for multiple testing correction in the work of (Aerts et al 2014, Hatt et al 2015, Yip et al 2016).

The optimal cutoff values of radiomic features are often used to stratify patients into two risk groups for Kaplan–Meier survival analysis (Cheng et al 2013a, Cook et al 2013). However, searching for the optimal cutoff values through testing multiple cutoffs can increase the likelihood of obtaining spurious significant results (Hilsenbeck et al 1992). Moreover, as the optimal cutoff value can vary in different datasets, the results may not be reproducible in different studies. Selection of an optimal cutoff for survival analysis is not recommended or must be accompanied by properly corrected significant values (p-values) (Altman et al 1994, Chalkidou et al 2015).

Numerous methods can be applied to reduce the number of radiomic features (Guyon et al 2003). The selected features can then be combined using various multivariate (classification) models to predict treatment outcome, tumour genetics, prognosis, metastatic potential, etc. In a study, Parmar et al (2014) investigated the prognostic values of 440 radiomic features using fourteen feature selection methods and twelve classification models in >460 lung cancer patients (Parmar et al 2015a). They found that the choice of the classification model could lead to variations in the predictive values of the radiomic features up to >30%, while choosing different feature selection methods only led to variations of about 6%. Furthermore, they identified feature selection methods and classification models that were stable to data perturbation while maintaining a decent performance for prediction of outcomes.

5. Summary

Here, we have reviewed applications and challenges of radiomics. Researchers have proposed to use radiomic features, which aim to quantify various tumour phenotypes on medical images, to describe this heterogeneity and furthermore, utilize these features as predictors of genetics and clinical outcomes. Despite the promising clinical potential of radiomics, there are precautions that must be taken in designing radiomics studies. For example, not all radiomics features are recommended for use due to their sensitivity to acquisition modes and reconstruction parameters. To examine the prognostic power of radiomic features, datasets consisting of ten to fifteen patients per feature evaluated has been recommended. Furthermore, the correlation of tumour volume and radiomic features should be reported to indicate the potential complementary value of the measures. Ideally, independent validation datasets are needed to confirm the prognostic value of the same radiomic features.

Acknowledgments

The authors would like to acknowledge the support from National Institute of Health (Award Number U01CA190234 and U24CA194354) and research seed funding grant from the American Association of Physicists in Medicine. The authors would also like to thank Dr Elizabeth Huynh for editorial assistance.
References

Aerts H J W L et al 2014 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach Nat. Commun. 5 4006


Balagurunathan Y et al 2014a Reproducibility and prognosis of quantitative features extracted from CT images Transl. Oncol. 7 72–87

Balagurunathan Y et al 2014b Test–retest reproducibility analysis of lung CT image features J. Digit. Imaging 27 805–23


Boellaard R 2011 Need for standardization of 18F-FDG PET/CT for treatment response assessments J. Nucl. Med. 52 93S–100S


Ella Naqa I et al 2009 Exploring feature-based approaches in PET images for predicting cancer treatment outcomes Pattern Recognit. 42 1162–71


Fanchon L M et al 2015 Feasibility of in situ, high-resolution correlation of tracer uptake with histopathology by quantitative autoradiography of biopsy specimens obtained under 18F-FDG PET/CT guidance J. Nucl. Med. 56 538–44


Galavis P E, Hollensen C, Jallow N, Paliwal B and Jeraj R 2010 Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters Acta Oncol. 49 1012–6


Ganeshan B, Abaleke S, Young R C D, Chatwin C R and Miles K A 2010 Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage Cancer Imaging 10 137–43


Maley C C et al 2006 Genetic clonal diversity predicts progression to esophageal adenocarcinoma Nat. Genet. 38 468–73


Naeini K M et al 2013 Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images Neuro-Oncology 15 626–34


Ng F, Kozarski R, Ganeshan B and Goh V 2013 Assessment of tumor heterogeneity by CT texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis? Eur. J. Radiol. 82 342–8


Soret M, Bacharach S L and Buvat I 2007 Partial-volume effect in PET tumor imaging J. Nucl. Med. 48 932–45


Tateishi U, Kusumoto M, Nishihara H, Nagashima K, Morikawa T and Moriyama N 2002 Contrast-enhanced dynamic computed tomography for the evaluation of tumor angiogenesis in patients with lung carcinoma Cancer 95 835–42


