

Bernard E. Van Beers, Benjamin Leporq,
Sabrina Doblas, and Philippe Garteiser

8.1 Introduction

Imaging biomarkers are health or disease markers based on quantitative imaging parameters. With high-throughput computing, it is now possible to extract numerous quantitative features from computed tomography (CT), magnetic resonance (MR), and positron emission tomography (PET) images. The conversion of digital medical images into mineable high-dimensional data is called radiomics and is motivated by the concept that biomedical images contain information that reflects underlying pathophysiology [1, 2]. The image measurements are based on size, volume, and shape assessment and on signal intensity and heterogeneity (texture) analysis.

8.2 Size Measurements

The simple clinically used metrics to assess lesion evolution include two-diameter (World Health Organization, WHO) and more recently, one-diameter (Response Evaluation Criteria in Solid

Tumors, RECIST) measurements [3–5]. For the last 15 years, the international cancer community has extensively employed the RECIST criteria at CT to assess the response exhibited by patient’s tumor on exposure to both marketed and experimental antitumor therapies [6]. The calculated response is categorized as complete response (disappearance of tumor), partial response (change between –100 and –30%), stable disease (change between –30 and +20%), or progressive disease (increase of 20% or greater). RECIST quantification of response correlates with patient survival and disease-free survival, showing its clinical usefulness [6].

However, RECIST criteria have several shortcomings. First, tumor evolution is linear, rather than polytomous. As cutoffs to define partial response or progressive disease are artificial, quantitative measurements are superior to semi-quantitative category assessment for studying tumor progression [6–9].

Second, the reproducibility of manual measurements may be suboptimal and may be improved by semiautomatic size measurements [10]. In a study of large lung tumors, it was shown that the 95% limits of inter-observer agreement (–39–28%) of maximum diameter measurements were outside the range of clinical acceptability (<20% according to the RECIST guidelines) at CT, whereas the corresponding automated measurements (–8.0–11%) were within clinical acceptable range [11].

B.E. Van Beers (✉) • B. Leporq • S. Doblas
P. Garteiser
Laboratory of Imaging Biomarkers, UMR 1149
INSERM – University Paris Diderot, Paris, France
Beaujon University Hospital Paris Nord,
Paris, France
e-mail: bernard.van-beers@aphp.fr

Third, RECIST size measurements do not always accurately reflect tumor response, especially when molecular therapies or other targeted therapeutic interventions such as chemoembolization are used [12, 13]. This is explained by the fact that these treatments mainly cause tumor necrosis, with little or no size decrease.

Alternative response criteria have been developed in these cases. These criteria include the mRECIST criteria, in which one diameter of the viable, contrast-enhancing, tumor regions is measured; the European Association for the Study of Liver Disease (EASL) criteria in which two diameters of the enhancing regions are measured; and the Choi criteria in which decrease in tumor size and decrease in tumor density at CT are assessed [14].

In patients with hepatocellular carcinomas, the Choi criteria have been shown to be superior to the RECIST, mRECIST, and EASL criteria to assess treatment response [15]. This underscores the fact that combining signal intensity measurements with size measurements may increase the diagnostic value relative to size measurements alone. With the Choi method, however, the signal attenuation measurements are obtained as the mean value within a region of interest. This region of interest analysis provides only part of the information as tumor heterogeneity is not explicitly described.

8.3 Lesion Segmentation

For more complete quantitative assessment of lesions, feature measurements within the whole lesion volume are needed. Three-dimensional volume segmentation is a critical and challenging component of whole lesion analysis. It is critical because subsequent parameters are generated from the segmented volumes. It is challenging because many tumors have indistinct borders.

Multiple segmentation algorithms have been applied in medical imaging studies. Popular ones are based on boundary or active contour definition [1, 16], region-growing or level-set methods [17, 18], and k-means clustering approaches [19, 20].

Active contours consist of positioning a contour larger than the region to be segmented and iteratively repositioning its points until a convergence criterion is met. The convergence criterion may be based on the geometry of the contour, thereby introducing prior knowledge on the shape of the segmented region [21], on the intensity and spatial variations thereof over the underlying region [22], or on a combination of both types of information. Region-growing approaches, and their advanced counterparts, namely, level-set methods, consist of starting an iterative process on an initial position for the region of interest. This region is then augmented or “grown,” by adding neighboring pixels to it. Addition of pixels is conditioned positively if the resulting, larger region remains homogeneous, and negatively if the homogeneity decreases, indicative of a boundary [23]. Finally, k-means clustering approaches rely on Euclidean measures of distances between extracted parameters (pixel intensity or other pixel-wise derived metrics) to generate pixel clusters corresponding to homogeneous regions [24].

Accuracy and reproducibility are important factors to evaluate segmentation algorithms for medical images. However, accuracy is difficult to determine because the reference method is often based on manual segmentation, which is subjective, error prone, and time consuming. Objective volume measurements during surgery are better gold standards but are rarely obtained [17]. In other words, “ground truth” segmentation often does not exist.

Hence, reproducibility is more important than accuracy. Several studies have shown that the reproducibility of semiautomatic segmentation algorithms is superior to that of manual segmentation [11, 17, 18, 25]. A consensus is emerging that optimum reproducible segmentation is achievable with computer-aided edge detection followed by manual curation [2].

8.4 Shape-Based Measurements

Quantitative features describing the geometric shape of a lesion can be extracted from the three-dimensional surface of the rendered volumes.

Measures of compactness, spherical disproportion, sphericity, surface-to-volume ratio, and Zernike moments describe the shape of the lesion [26–28].

8.5 Intensity and Texture Analyses

Intensity and texture analyses can be divided into four families based on the distribution of signal intensity, on the organization of gray level in the spatial domain, on the organization of geometric patterns in the spatial domain, and on analysis performed in the frequency domain.

8.5.1 Analysis Based on the Distribution of Signal Intensity

This analysis is based on first-order statistics which describe the distribution of values of individual voxels without concern for spatial relationships. These are generally histogram-based methods and reduce a region of interest to single values. The parameters include the mean, median, maximum and minimum values, *n*th centiles, standard deviation, variance, mean absolute deviation, uniformity (uniformity of gray-level distribution), entropy (irregularity of gray-level distribution), skewness (asymmetry of the histogram), and kurtosis.

8.5.2 Analysis Based on the Organization of Signal Intensity in the Image Domain

This analysis provides second-order descriptors which describe statistical interrelationships between voxels with similar or dissimilar contrast values. The spatial distribution of voxel intensities is calculated from gray-level co-occurrence (GLCM) or gray-level run-length texture matrices (GLRLM).

GLCM determines how often a pixel of intensity *i* finds itself within a certain relationship to

another pixel of intensity *j* (Fig. 8.1). Second-order statistics based on a co-occurrence matrix (GLCM) include autocorrelation, contrast, correlation, cluster prominence, cluster shade, cluster tendency, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares, sum average, sum variance, sum entropy, etc. [29]. The energy (pixel repetition) expresses the regularity of the texture. High energy is observed when the high values in the GLCM are concentrated in some precise locations. It is the case for images with constant or periodic gray-level distributions. A random or noisy image gives a GLCM with more distributed values and a low energy. The contrast is more elevated for GLCM with larger values outside the diagonal, thus for images with local variation of intensities.

The dissimilarity expresses the same characteristic than the contrast, but the weights of inputs of the GLCM increase linearly from the diagonal rather than quadratically for the contrast. These two descriptors are thus often correlated.

The entropy (randomness of the matrix) relies to the spreading of the GLCM diagonal. The entropy is the inverse of energy. These parameters are often correlated.

The homogeneity (uniformity of co-occurrence matrix) inversely evolves with the contrast. Homogeneity is high when the differences between co-occurrences are small. It is more sensitive on the diagonal elements of the GLCM than the contrast which depends on elements outside the diagonal.

The correlation may be described as a measurement of the linear dependency of gray levels of the image. The cluster shade and cluster prominence give information about the degree of symmetry of the GLCM. High values represent low symmetric pattern.

The main difficulty when using GLCM is to fix the parameters because this step needs to be performed case by case. The distance *d* must reflect the local correlation between the pixels. It is admitted that the correlation is more pertinent for short distances and, typically, *d* is fixed equal to 1. In practice, GLCM is computed over four orientations (i.e., 0°, 45°, 90°, and 135°) according to Haralick recommendations [29]. The

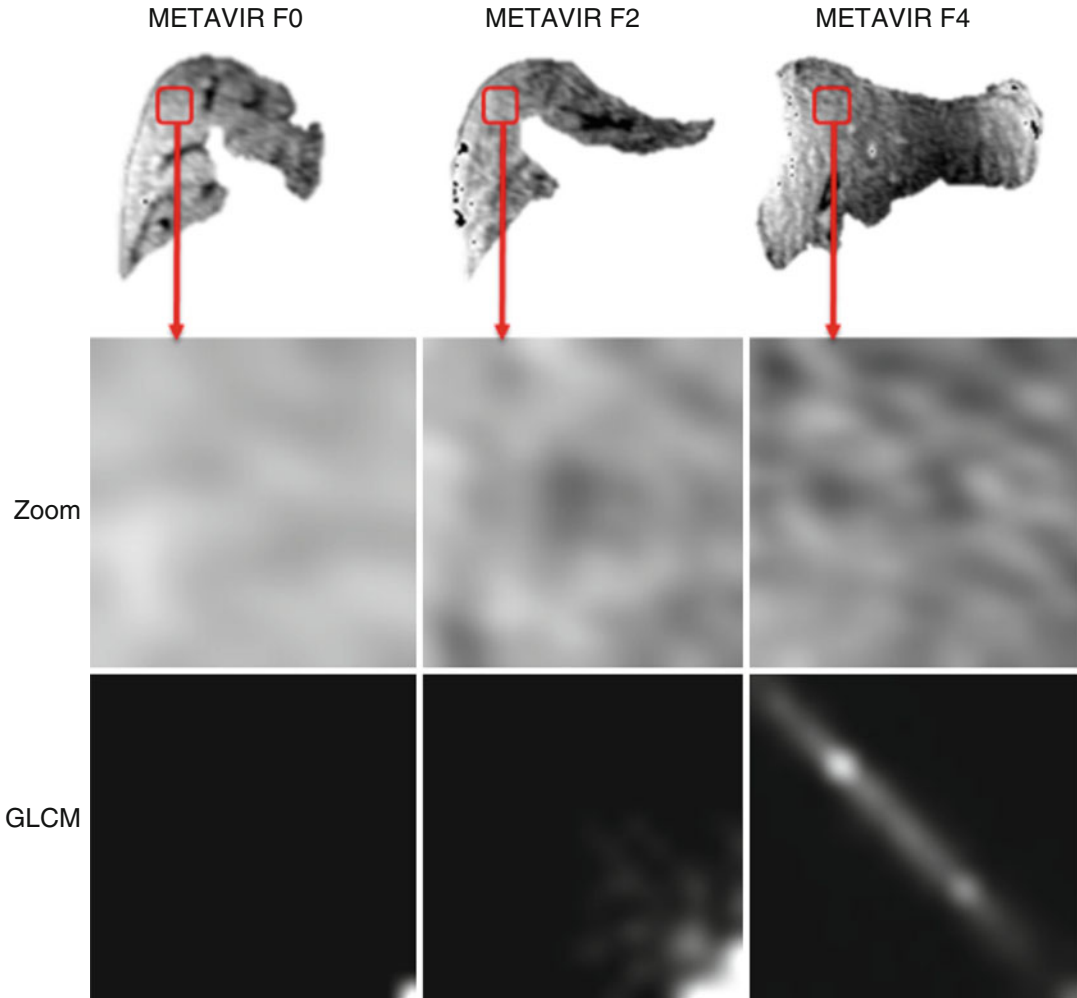


Fig. 8.1 Texture analysis of gadoxetic acid-enhanced MR images in patient with chronic liver disease. This figure shows differences in the GLCM according to the

severity of liver fibrosis. Second-order descriptors derived from this matrix can offer quantitative information relevant for the assessment of liver fibrosis

features are computed for each orientation and can be concatenated in a single array of descriptors or averaged to obtain an array of descriptors invariant regarding to the rotation. The choice of the window (i.e., the number of gray levels in the parametric image) is also important and imposes a compromise between the pertinence of the descriptors and the fidelity of the texture.

Another method to derive second-order statistics is the gray-level run-length matrix (GLRLM). A gray-level run is defined as the length in number of pixels of consecutive pixels that have the same gray-level value. From the GLRLM,

features can be extracted describing short- and long-run emphasis, gray-level nonuniformity, run-length nonuniformity, run percentage, low gray-level run emphasis, and high gray-level run emphasis [1, 28]. The short-run emphasis characterizes the smoothness of the texture, whereas the long-run emphasis characterizes the coarseness. The run percentage is the ratio between the number of runs over the number of pixels in the image. It characterizes the homogeneity of the texture. The gray-level nonuniformity measures the uniformity of run distribution. It is minimal when the runs are uniformly distributed between

the gray levels. The run-length nonuniformity measures the uniformity of run length and increases with the number of runs of same length.

Other matrices have been proposed to characterize the texture in the spatial domain such as the gray-level size zone matrix (GLSZM). GLSZM does not require computation in several directions, in contrast to GLRLM and GLCM. However, the degree of gray-level quantization has an important impact on the texture classification performance. Similarly to GLRLM, descriptors can be derived from the analysis of this matrix such as the small-zone size emphasis, large-zone size emphasis, low gray-level zone emphasis, high gray-level zone emphasis, small-zone low-gray emphasis, small-zone high-gray emphasis, large-zone low-gray emphasis, large-zone high-gray emphasis, gray-level nonuniformity, zone size nonuniformity, and zone size percentage [30].

8.5.3 Analysis Based on the Organization of Geometric Pattern in the Image Domain

Filter grids can be applied on the images to extract repetitive or non-repetitive patterns. These methods include fractal analysis, wherein patterns are imposed on the images and the number of grid elements containing voxels of a specified value is computed; Minkowski functionals, which assess patterns of voxels whose intensity is above a threshold [31]; and Laplacian transforms of Gaussian band-pass filters that extract areas with increasingly coarse texture patterns from the images [32].

8.5.4 Texture Analysis in the Frequency Domain

These methods use filtering tools such as the Fourier transform, the wavelet decomposition, and the Gabor filter to extract the information. The 2D Fourier transform allows to represent the frequency spectrum on images in which each coefficient corresponds to a frequency in a given

orientation. Therefore, the center of the spectra includes the low frequencies and the extremities the high frequencies. An image with a smooth texture will display a spectrum with high values concentrated close to the center, whereas an image with a rough texture will display a spectra with high value concentrated at the extremities. Quantitative information related to the texture can be extracted by decomposing the spectra into sub-bands according to their polar coordinates and calculating the average, energy, variance, and maximum [33]. The Fourier transform can also be applied in local neighbors in the image. It is possible to determine a radial spectrum on windows with increasing size by averaging the coefficient of the Fourier spectrum over all orientations. A principal component analysis is performed to identify the range of frequencies and the size window explaining the variability [34]. The Fourier spectrum only contains frequency information.

In contrast, Gabor filters and the wavelet transforms provide both frequency and spatial information. Gabor filters have the ability to model the direction and frequency sensitivity by decomposing the image spectrum in a narrow range of frequencies and orientations. In the spatial domain, the Gabor filter is a Gaussian function modulated by a complex sinusoid and a Gaussian surface centered on a central frequency F with an orientation θ in the frequency domain. A conventional practice with Gabor filters consists in using filter banks, each centered on a different central frequency and orientation, by covering the whole frequency domain. Each pixel gives a response for each filter. To have a different proportion covered by each filter and to limit the overlap, thus the redundancy of information, Manjunath and Ma have proposed to decompose the spectrum in several scales and orientations [35]. Mean and standard deviation of the filter responses are calculated to extract the texture signature.

Nevertheless, due to the non-orthogonality of Gabor filters, texture attributes derived from these filters can be correlated. It is difficult to determine if a similarity observed between the analysis scales is linked to the property of the

image or to redundancy in the information. Thus for each scale of application, parameters defining the filter must be modified.

This issue is addressed by the use of wavelets, offering a uniform analysis framework by decomposing the image into orthogonal and independent sub-bands. Briefly, the wavelet decomposes the image with a series of functions obtained by translation and scaling from an initial function, called mother wavelet. Wavelet decomposition of an image is the convolution product between the image and the wavelet functions [31].

8.6 Data Reduction

The number of descriptive image features can approach the complexity of data obtained with gene expression profiling. With such large complexity, there is a danger of overfitting analyses, and hence, dimensionality must be reduced by prioritizing the features. Dimensionality reduction can be divided into feature extraction and feature selection. Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions, as in principal component analysis.

Feature selection techniques can be broadly grouped into approaches that are classifier dependent (wrapper and embedded methods) and classifier independent (filter methods). Wrapper methods search the space of feature subsets, using the training/validation accuracy of a particular classifier as the measure of utility for a candidate subset. This may deliver significant advantages in generalization, though has the disadvantage of a considerable computational expense, and may produce subsets that are overly specific to the classifier used. As a result, any change in the learning model is likely to render the feature set suboptimal. Embedded methods exploit the structure of specific classes of learning models to guide the feature selection process. These methods are less computationally expensive, and less prone to overfitting than wrappers, but still use quite strict model structure assumptions.

In contrast, filter methods evaluate statistics of the data independently of any particular classifier, thereby extracting features that are generic, having incorporated few assumptions. Each of these three approaches has its advantages and disadvantages, the primary distinguishing factors being speed of computation, and the chance of overfitting. In general, in terms of speed, filters are faster than embedded methods which are in turn faster than wrappers. In terms of overfitting, wrappers have higher learning capacity so are more likely to overfit than embedded methods, which in turn are more likely to overfit than filter methods.

A primary advantage of filters is that they are relatively cheap in terms of computational expense and are generally more amenable to a theoretical analysis of their design. The defining component of a filter method is the relevance index quantifying the utility of including a particular feature in the set. The filter-based feature selection methods can be divided into two categories: univariate methods and multivariate methods. In case of univariate methods, the scoring criterion only considers the relevancy of features ignoring the feature redundancy, whereas multivariate methods investigate the multivariate interaction within features, and the scoring criterion is a weighted sum of feature relevancy and redundancy [36–38].

One of the simplest methods relies on the computation of cross correlation matrices, whereby the correlation between each pair of features is computed (Fig. 8.2). The resulting matrix is subsequently thresholded to identify subsets of features that are highly correlated.

A single feature from each subset can then be selected based on maximum relevancy.

8.7 Data Classification

For data mining, unsupervised and supervised analysis options are available. The distinction in these approaches is that unsupervised analysis does not use any outcome variable, but rather provides summary information and graphical representations of the data. Supervised analysis,

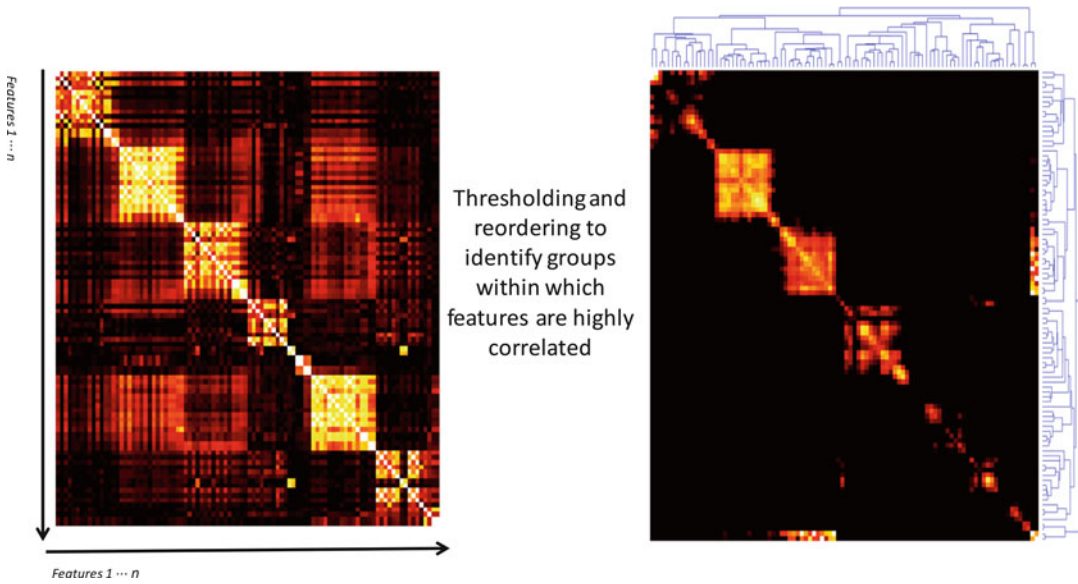


Fig. 8.2 Illustration of the feature selection process. The cross correlation matrix on the *left* is reordered with linkage algorithms on the *right* and thresholded to a given value of

correlation coefficient. For data analysis, one feature of each group on the right can be selected based on maximum relevancy, e.g., maximum patient interpatient variability

in contrast, creates models that attempt to separate or predict the data with respect to an outcome or phenotype.

Clustering is the grouping of like data and is one of the most common unsupervised analysis approaches. There are many different types of clustering. Hierarchical clustering, or the assignment of examples into clusters at different levels of similarity into a hierarchy of clusters, is a common type. Similarity is based on correlation (or Euclidean distance) between individual examples or clusters.

Alternatively, k-means clustering is based on minimizing the clustering error criterion which for each point computes its squared distance from the corresponding cluster center and then takes the sum of these distances for all points in the data set.

The data from this type of analyses can be graphically represented using the cluster color map. Cluster relationships are indicated by tree-like structures adjacent to the color map or by k-means cluster groups [24, 39] (Fig. 8.3).

Supervised analysis consists of building a mathematical model of an outcome or response variable. The breadth of techniques available is

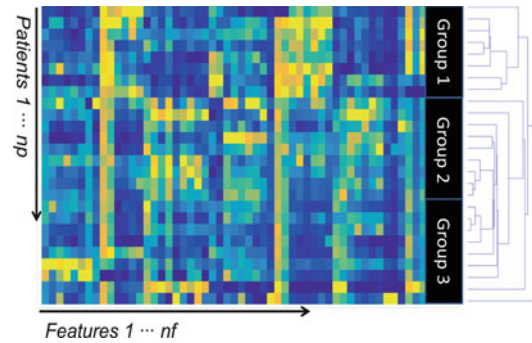


Fig. 8.3 Graphical representation of a radiomics data set. Each patient represents a row of the matrix (np , number of patients), and each column represents one of the features (nf , number of features). First-order imaging parameters based on MR elastography data acquired at several mechanical frequencies in patients with liver cirrhosis and portal hypertension. The hierarchical cluster relationships are indicated by tree-like structures on the right of the matrix representation. Alternatively, clustering by a k-means algorithm can be used to group patients into like groups, indicated by groups one to three in the black boxes

remarkable and includes neural networks, decision trees, classification, and regression trees as well as Bayesian networks [40, 41]. Model selection is dependent on the nature of the outcome and the nature of the training data.

Performance in the training data set is always upward biased because the features were selected from the training data set. Therefore, a validation data set is essential to establish the likely performance in the clinic. Preferably, validation data should come from an external independent institution or trial [41]. Alternatively, one may evaluate machine learning algorithms on a particular data set, by partitioning the data set in different ways. Popular partition strategies include k-fold cross validation, leave-one-out, and random sampling [42].

The best models are those that are tailored to a specific medical context and, hence, start out with a well-defined end point. Robust models accommodate patient features beyond imaging. Covariates include genomic profiles, histology, serum biomarkers, and patient characteristics [2].

As a general rule, several models should be evaluated to ascertain which model is optimal for the available data [38, 43]. Recently, Ypsilantis et al. [44] have compared the performance of two competing radiomics strategies: an approach based on state-of-the-art statistical classifiers (logistic regression, gradient boosting, random forests, and support vector machines) using over 100 quantitative imaging descriptors, including texture features as well as standardized uptake values and a convolutional neural network, trained directly from PET scans by taking sets of adjacent intra-tumor slices. The study was performed for predicting response to neoadjuvant chemotherapy in patients with esophageal cancer, from a single 18F-FDG-PET scan taken prior to treatment. The limitation of the statistical classifiers originates from the fact that the performance is highly dependent on the design of the texture features, thus requiring prior knowledge for a specific task and expertise in hand-engineering the necessary features. By contrast, convolutional neural networks operate directly on raw images and attempt to automatically extract highly expressive imaging features relevant to a specific task at hand. In the Ypsilantis et al. study, convolutional neural networks achieved 81 % sensitivity and 82 % specificity in predicting nonresponders and outperformed the

other competing predictive models. These results suggest the potential superiority of the fully automated method. However, further testing using larger data sets is required to validate the predictive power of convolutional neural networks for clinical decision-making.

Indeed, it should be noted that machine learning techniques in radiology are still in infancy. Many machine learning studies were done using relatively small data sets. The proposed methods may not generalize well from small data sets to large data sets. To solve the problem, re-training the algorithm will be necessary, but it requires intervention of knowledgeable experts which hinders the deployment of machine learning-based systems in hospitals or medical centers. One possible solution would be utilizing incremental learning and adjusting the computerized systems in an automatic way. In addition, increased large-scale data may bring computational issues to radiology applications. Machine learning techniques employed in these applications may not scale well as training data increases [42].

8.8 Radiomics

Radiomics mines and deciphers numerous medical imaging features. The hypothesis being that these imaging features are augmented with critical and interchangeable information regarding tumor phenotype [28]. Texture is especially important to assess in tumors. Indeed, the tumor signal intensity is very heterogeneous and reflects its structural and functional features, including the number of tumor cells, quantity of inflammation and fibrosis, perfusion, diffusion, and mechanical properties, as well as metabolic activity. Functional parameters which are hallmarks of cancer include sustaining proliferative signaling, resisting cell death, inducing angiogenesis, activating invasion and metastasis, and deregulating cellular energetics [45]. These hallmarks can be assessed with quantitative MR imaging, including perfusion and diffusion MR imaging, MR elastography and susceptibility, and FDG-PET [46, 47].

During the recent years, it became increasingly evident that genetic heterogeneity is a basic feature of cancer and is linked to cancer evolution [48]. This heterogeneity which evolves during time concerns not only the tumor cells but also their microenvironment [49]. Moreover, it has been shown that the global gene expression patterns of human cancers may systematically correlate with their dynamic imaging features [50]. Tumors are thus characterized by regions habitats with specific combinations of blood flow, cell density, necrosis, and edema. Clinical imaging is uniquely suited to measure temporal and spatial heterogeneity within tumors [51], and this information may have predictive and prognostic value.

Spatial heterogeneity is found between different tumors within individual patients (inter-tumor heterogeneity) and within each lesion in an individual (intra-tumor heterogeneity). Intra-tumor heterogeneity is near ubiquitous in malignant tumors, but the extent varies between patients. Intra-tumor heterogeneity tends to increase as tumors grow. Moreover, established spatial heterogeneity frequently indicates poor clinical prognosis. Finally, intra-tumor heterogeneity may increase or decrease following efficacious anticancer therapy, depending on underlying tumor biology [52].

Several studies have shown that tumor heterogeneity at imaging may predict patient survival or response for treatment [53–59].

For instance, in 41 patients with newly diagnosed esophageal cancer treated with combined radiochemotherapy, Tixier et al. showed that textural features of tumor metabolic distribution extracted from baseline ^{18}F -FDG-PET images allowed for better prediction of therapy response than first-order statistical outputs (mean, peak, and maximum SUV) [60].

In 26 colorectal cancer liver metastases, O'Connor et al. showed that three perfusion parameters, namely, the median extravascular extracellular volume, the heterogeneity parameters corresponding to tumor-enhancing fraction, and the microvascular uniformity (assessed with the fractal measure box dimension), explained 86% of the variance tumor shrinkage after FOLFOX therapy [61]. This underscores that

measuring microvascular heterogeneity may yield important prognostic and/or predictive biomarkers.

Zhou et al. showed in 32 patients with glioblastoma multiforme that spatial variations in T1 post-gadolinium and either T2-weighted or fluid-attenuated inversion recovery at baseline MR imaging correlated significantly with patient survival [62].

8.9 Limitations of Radiomics

Several issues arise when interpreting imaging data of heterogeneity. First, some voxels suffer from partial volume averaging, typically at interface with non-tumor 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. Most methods of analysis require hundreds to thousands of voxels for robust application. Third, CT, MR imaging, or PET voxels are usually non-isotropic (slice thickness exceeds in-plane resolution). Dimensions are typically 200–2,000 μm for rodent models and 500–5,000 μm for clinical tumors. Compared with genomic and histopathology biomarkers, this represents many orders of magnitude difference in scale, making it difficult to validate image heterogeneity biomarkers against pathology [52].

Variations in image parameters affect the information being extracted by image feature algorithms, which in turn affects classifier performance (Fig. 8.4) [63]. At PET imaging, Yan et al. [64] analyzed the effect of several acquisition parameters on the heterogeneity values. They found that the voxel size affected the heterogeneity value the most, followed by the full width at half maximum of the Gaussian post-processing filter applied to the reconstructed images. Neither the number of iterations nor the actual reconstruction scheme affected the heterogeneity values much.

Because of the information dependence on variations in image parameters, imaging standardization and reproducibility are important issues to determine the effectiveness of image

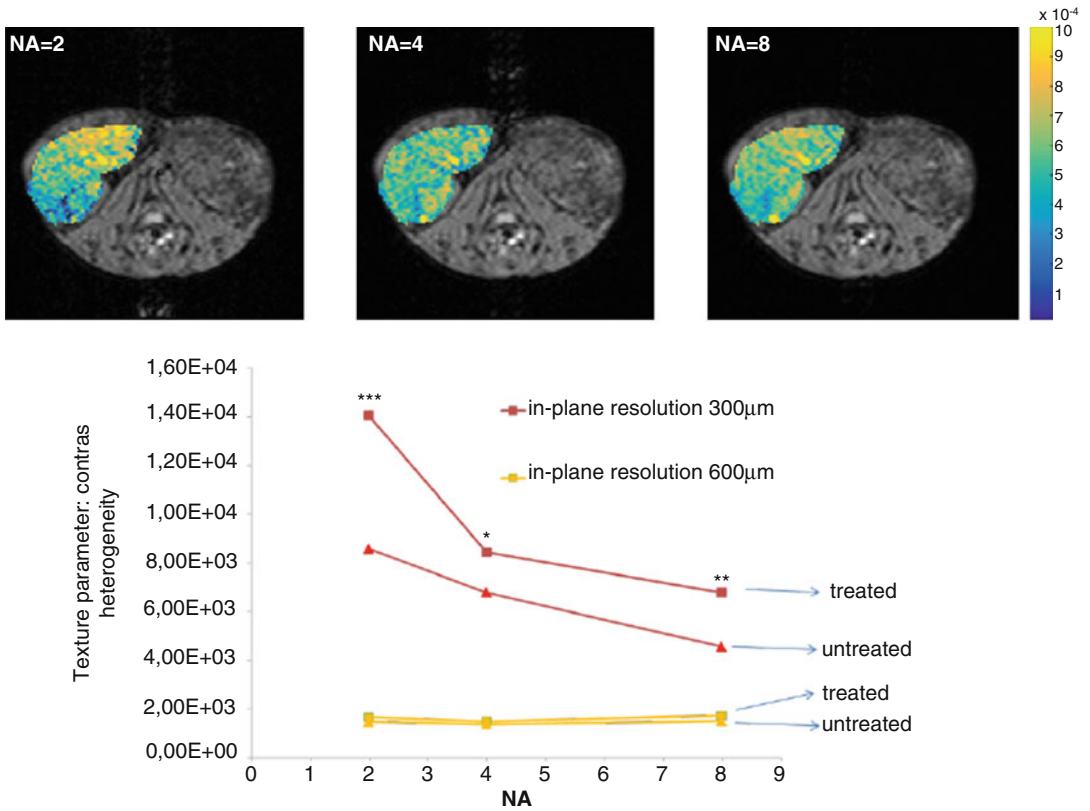


Fig. 8.4 The behavior of a texture parameter acquired on diffusion coefficient maps was assessed in a HepG2 tumor treated with an adipokine that competitively inhibits the fatty acid-binding protein Fabp4. The effects of in-plane resolution and number of averages, were explored. The treated tumor had significantly higher texture on the high-

resolution data set regardless of signal-to-noise ratio. On the low-resolution data sets, adipokine treatment did not appear to have an effect. These data show that spatial resolution and signal to noise ratios (manipulated here through varying number of averages (NA)) may affect texture analysis

features being developed and prediction models built to work on those feature values.

Another problem in radiomics and genomics is related to multiple testing issues. In many data sets in these areas, it is not unusual to test the significance of thousands of variables using 50 samples. Any single test may have a low expected false-positive rate; however, the cumulative effect of many repeated tests guarantees that many statistically significant findings are due to random chance (type I errors in statistics should be <5%). Chalkidou et al. reported a systematic review of the type I error inflation in texture analysis derived from PET or CT images [65]. After applying appropriate statistical corrections, an average type I error probability of 76%

was estimated with the majority of published results not reaching statistical significance. This underscores that the multiple testing problem may be critical. It has been addressed in statistics in many ways. However, the best way to overcome overfitting and optimism in predictive performance is to evaluate the performance of the model in an external validation cohort, as explained above [66].

Conclusion

Current knowledge suggests that radiomics can enhance individualized treatment selection and monitoring. Furthermore, unlike genomics-based approaches, radiomics is noninvasive and comparatively cost-effective.

Radiomics is thus an innovative and encouraging breakthrough toward the realization of precision medicine. Fast-computing and state-of-the-art software have facilitated the collection and analysis of large amounts of data, while the development of data mining techniques enables researchers to test a large number of hypotheses simultaneously. The high number of image analysis algorithms and image-derived features is promising to unravel complex biology by overcoming the limitations inherent in invasive tissue sampling techniques. However, the high data dimensionality complicates the quantitative analysis, and robust biological and statistical validation is needed before advanced radiomics solutions can be used in the clinics.

References

1. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234–48.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures they are data. *Radiology*. 2016;278(2):563–77. Epub 2015/11/19.
3. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer*. 1981;47(1):207–14. Epub 1981/01/01.
4. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst*. 2000;92(3):205–16. Epub 2000/02/03.
5. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–47. Epub 2008/12/23.
6. Jain RK, Lee JJ, Ng C, Hong D, Gong J, Naing A, et al. Change in tumor size by RECIST correlates linearly with overall survival in phase I oncology studies. *J Clin Oncol*. 2012;30(21):2684–90. Epub 2012/06/13.
7. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J Natl Cancer Inst*. 2007;99(19):1455–61. Epub 2007/09/27.
8. Raymond E, Dahan L, Raoul JL, Bang YJ, Borbath I, Lombard-Bohas C, et al. Sunitinib malate for the treatment of pancreatic neuroendocrine tumors. *N Engl J Med*. 2011;364(6):501–13. Epub 2011/02/11.
9. Sharma MR, Maitland ML, Ratain MJ. RECIST: no longer the sharpest tool in the oncology clinical trials toolbox--point. *Cancer Res*. 2012;72(20):5145–9; discussion 50. Epub 2012/09/07.
10. Bonekamp D, Bonekamp S, Halappa VG, Geschwind JF, Eng J, Corona-Villalobos CP, et al. Interobserver agreement of semi-automated and manual measurements of functional MRI metrics of treatment response in hepatocellular carcinoma. *Eur J Radiol*. 2014;83(3):487–96. Epub 2014/01/07.
11. Dinkel J, Khalilzadeh O, Hintze C, Fabel M, Puderbach M, Eichinger M, et al. Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. *Lung Cancer*. 2013;82(1):76–82. Epub 2013/08/13.
12. Le Cesne A, Van Glabbeke M, Verweij J, Casali PG, Findlay M, Reichardt P, et al. Absence of progression as assessed by response evaluation criteria in solid tumors predicts survival in advanced GI stromal tumors treated with imatinib mesylate: the intergroup EORTC-ISG-AGITG phase III trial. *J Clin Oncol*. 2009;27(24):3969–74. Epub 2009/07/22.
13. Lencioni R, Lovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. *Semin Liver Dis*. 2010;30(1):52–60. Epub 2010/02/23.
14. Choi H, Chamsangavej C, Faria SC, Macapinlac HA, Burgess MA, Patel SR, et al. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: proposal of new computed tomography response criteria. *J Clin Oncol*. 2007;25(13):1753–9. Epub 2007/05/02.
15. Ronot M, Bouattour M, Wassermann J, Bruno O, Dreyer C, Larroque B, et al. Alternative Response Criteria (Choi, European association for the study of the liver, and modified Response Evaluation Criteria in Solid Tumors [RECIST]) Versus RECIST 1.1 in patients with advanced hepatocellular carcinoma treated with sorafenib. *Oncologist*. 2014;19(4):394–402. Epub 2014/03/22.
16. Michoux N, Vallee JP, Pechere-Bertschi A, Montet X, Buehler L, Van Beers BE. Analysis of contrast-enhanced MR images to assess renal function. *MAGMA*. 2006;19(4):167–79. Epub 2006/08/15.
17. Hermoye L, Laamari-Azjal I, Cao Z, Annet L, Lerut J, Dawant BM, et al. Liver segmentation in living liver transplant donors: comparison of semiautomatic and manual methods. *Radiology*. 2005;234(1):171–8. Epub 2004/11/27.
18. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One*. 2014;9(7), e102107. Epub 2014/07/16.

19. Michoux N, Simoni P, Tombal B, Peeters F, Machiels JP, Lecouvet F. Evaluation of DCE-MRI postprocessing techniques to assess metastatic bone marrow in patients with prostate cancer. *Clin Imaging*. 2012;36(4):308–15. Epub 2012/06/26.
20. Michoux N, Van den Broeck S, Lacoste L, Fellah L, Galant C, Berliere M, et al. Texture analysis on MR images helps predicting non-response to NAC in breast cancer. *BMC Cancer*. 2015;15:574. Epub 2015/08/06.
21. Pathak SD, Chalana V, Kim Y. Interactive automatic fetal head measurements from ultrasound images using multimedia computer technology. *Ultrasound Med Biol*. 1997;23(5):665–73. Epub 1997/01/01.
22. Sebbahi A, Herment A, de Cesare A, Mousseaux E. Multimodality cardiovascular image segmentation using a deformable contour model. *Comput Med Imaging Graph*. 1997;21(2):79–89. Epub 1997/03/01.
23. Hojjatoleslami SA, Kittler J. Region growing: a new approach. *IEEE Trans Image Process*. 1998;7(7):1079–84. Epub 2008/02/16.
24. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. *Pattern Recogn*. 2003;36:451–61.
25. Heye T, Merkle EM, Reiner CS, Davenport MS, Horvath JJ, Feuerlein S, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part II. Comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. *Radiology*. 2013;266(3):812–21. Epub 2012/12/12.
26. Tahmasbi A, Saki F, Shokouhi SB. Classification of benign and malignant masses based on Zernike moments. *Comput Biol Med*. 2011;41(8):726–35. Epub 2011/07/05.
27. Yap FY, Bui JT, Knuttinen MG, Walzer NM, Cotler SJ, Owens CA, et al. Quantitative morphometric analysis of hepatocellular carcinoma: development of a programed algorithm and preliminary application. *Diagn Interv Radiol*. 2013;19(2):97–105. Epub 2012/12/13.
28. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. Epub 2014/06/04.
29. Haralick RMSK, Dinstein IH. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3(6):610–21.
30. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Trans Bio-Med Eng*. 2014;61(3):630–7. Epub 2013/10/11.
31. Alberich-Bayarri A, Marti-Bonmati L, Angeles Perez M, Sanz-Requena R, Lerma-Garrido JJ, Garcia-Marti G, et al. Assessment of 2D and 3D fractal dimension measurements of trabecular bone from high-spatial resolution magnetic resonance images at 3 T. *Med Phys*. 2010;37(9):4930–7. Epub 2010/10/23.
32. Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging*. 2012;3(6):573–89. Epub 2012/10/25.
33. Augusteijn MF, Clements L, Shaw KA. Performance evaluation of texture measures for ground cover identification in satellite images by means of a neural network classifier. *IEEE Trans Geosci Remote Sens*. 1995;33(3):616–26.
34. Proisy C, Couteron P, Fromard F. Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of IKONOS images. *Remote Sens Environ*. 2007;109:379–92.
35. Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Mach Intell*. 1996;18(8):837–42.
36. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. 2005;27(8):1226–38. Epub 2005/08/27.
37. Brown G, Pocock A, Zhao MJ, Lujan M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J Mach Learn Res*. 2012;13(1):27–66.
38. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol*. 2015;5:272. Epub 2015/12/24.
39. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95(25):14863–8. Epub 1998/12/09.
40. Lee SM, Abbott PA. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *J Biomed Inform*. 2003;36(4-5):389–99. Epub 2003/12/04.
41. Lambin P, Zindler J, Vanneste BG, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. *Adv Drug Deliv Rev*. 2016. pii:S0169-409X(16)30008-4. Epub 2016/01/18.
42. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal*. 2012;16(5):933–51. Epub 2012/04/03.
43. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep*. 2015;5:13087. Epub 2015/08/19.
44. Ypsilantis PP, Siddique M, Sohn HM, Davies A, Cook G, Goh V, et al. Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. *PLoS One*. 2015;10(9), e0137036. Epub 2015/09/12.
45. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–74. Epub 2011/03/08.
46. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for

- PET response criteria in solid tumors. *J Nucl Med.* 2009;50 Suppl 1:122S–50. Epub 2009/06/24.
47. Van Beers BE, Daire JL, Garteiser P. New imaging techniques for liver diseases. *J Hepatol.* 2015;62(3):690–700. Epub 2014/12/03.
 48. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature.* 2013;501(7467):338–45. Epub 2013/09/21.
 49. Junttila MR, de Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature.* 2013;501(7467):346–54. Epub 2013/09/21.
 50. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol.* 2007;25(6):675–80. Epub 2007/05/23.
 51. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. *Radiology.* 2013;269(1):8–15. Epub 2013/09/26.
 52. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res.* 2015;21(2):249–57. Epub 2014/11/26.
 53. Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, Liang Y, et al. Identification of non-invasive imaging surrogates for brain tumor gene-expression modules. *Proc Natl Acad Sci U S A.* 2008;105(13):5213–8. Epub 2008/03/26.
 54. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. *Radiology.* 2012;264(2):387–96. Epub 2012/06/23.
 55. Ng F, Ganeshan B, Kozarski R, Miles KA, Goh V. Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival. *Radiology.* 2013;266(1):177–84. Epub 2012/11/16.
 56. Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med.* 2015;56(1):38–44. Epub 2014/12/17.
 57. Cook GJ, Yip C, Siddique M, Goh V, Chicklore S, Roy A, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? *J Nucl Med.* 2013;54(1):19–26. Epub 2012/12/04.
 58. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol.* 2015;114(3):345–50. Epub 2015/03/10.
 59. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc Natl Acad Sci U S A.* 2015;112(46):E6265–73. Epub 2015/11/19.
 60. Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52(3):369–78. Epub 2011/02/16.
 61. O'Connor JP, Rose CJ, Jackson A, Watson Y, Cheung S, Maders F, et al. DCE-MRI biomarkers of tumour heterogeneity predict CRC liver metastasis shrinkage following bevacizumab and FOLFOX-6. *Br J Cancer.* 2011;105(1):139–45. Epub 2011/06/16.
 62. Zhou M, Hall L, Goldof D, Russo R, Balagurunathan Y, Gillies R, et al. Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results. *Transl Oncol.* 2014;7(1):5–13. Epub 2014/04/29.
 63. Buvat I, Orhac F, Soussan M. Tumor texture analysis in PET: where do we stand? *J Nucl Med.* 2015;56(11):1642–4. Epub 2015/08/22.
 64. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. *J Nucl Med.* 2015;56(11):1667–73. Epub 2015/08/01.
 65. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One.* 2015;10(5), e0124165. Epub 2015/05/06.
 66. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg.* 2015;102(3):148–58. Epub 2015/01/30.