HHS Public Access

Author manuscript

Eur J Nucl Med Mol Imaging. Author manuscript; available in PMC 2017 January 31.

Published in final edited form as:

Eur J Nucl Med Mol Imaging. 2017 January ; 44(1): 151–165. doi:10.1007/s00259-016-3427-0.

Characterization of PET/CT images using texture analysis: the past, the present… any future?

Mathieu Hatt[1], **Florent Tixier**[2,3], **Larry Pierce**[4], **Paul E. Kinahan**[4], **Catherine Cheze Le Rest**[2,3], and **Dimitris Visvikis**[1]

[1]INSERM, UMR 1101, LaTIM, University of Brest IBSAM, Brest, France

[2]Nuclear Medicine, University Hospital, Poitiers, France

[3]Medical school, EE DACTIM, University of Poitiers, Poitiers, France

[4]Imaging Research Laboratory, University of Washington, Seattle, WA, USA

Abstract

After seminal papers over the period 2009 – 2011, the use of texture analysis of PET/CT images for quantification of intratumour uptake heterogeneity has received increasing attention in the last 4 years. Results are difficult to compare due to the heterogeneity of studies and lack of standardization. There are also numerous challenges to address. In this review we provide critical insights into the recent development of texture analysis for quantifying the heterogeneity in PET/CT images, identify issues and challenges, and offer recommendations for the use of texture analysis in clinical research. Numerous potentially confounding issues have been identified, related to the complex workflow for the calculation of textural features, and the dependency of features on various factors such as acquisition, image reconstruction, preprocessing, functional volume segmentation, and methods of establishing and quantifying correspondences with genomic and clinical metrics of interest. A lack of understanding of what the features may represent in terms of the underlying pathophysiological processes and the variability of technical implementation practices makes comparing results in the literature challenging, if not impossible. Since progress as a field requires pooling results, there is an urgent need for standardization and recommendations/guidelines to enable the field to move forward. We provide a list of correct formulae for usual features and recommendations regarding implementation. Studies on larger cohorts with robust statistical analysis and machine learning approaches are promising directions to evaluate the potential of this approach.

Keywords

PET/CT; Image texture; Heterogeneity; Critical review; Recommendations

Correspondence to: Mathieu Hatt.

Electronic supplementary material The online version of this article (doi:10.1007/s00259-016-3427-0) contains supplementary material, which is available to authorized users.

Compliance with ethical standards

Conflicts of interest P.E.K. has received a research grant from GE Healthcare and is cofounder of PET/X LLC. The other authors declare that they have no conflicts of interest.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Introduction

Tumours are heterogeneous entities at all scales (macroscopic, physiological, microscopic, genetic) [1]. As a multimodal imaging modality, PET/CT is a promising tool for noninvasive exploration of intratumour heterogeneity at the macroscopic scale in both the anatomical and functional dimensions [2, 3]. The term heterogeneity usually conveys different meanings depending on the image modality. When considering the PET component, it refers to radiotracer uptake spatial distribution, which may reflect, depending on the radiotracer used, the combination of underlying biological processes such as metabolism, hypoxia, cellular proliferation, vascularization and necrosis [4–6]. Regarding the low-dose CT component of PET/CT, usually without contrast enhancement, heterogeneity refers to the variability in tissue density, which may result from spatially varying vascularization, necrosis or cellularity, as well as the proportions of fat, air and water [7]. With other modalities such as contrast-enhanced CT, as well as in MRI using various sequences (for example, T1, T2, FLAIR, DCE-MRI), heterogeneity can also include the spatial variability of vessel density, perfusion, proton density and physiological tissue characteristics [8–12].

The heterogeneity of image voxel intensities can be quantified by different image processing and analysis methods, including texture analysis (TA) [3], fractal analysis [13], shape models [14–16], intensity histogram analysis [15, 17] and filtering combined with statistical and frequency-based methods [18]. This critical review focuses on the use of TA in PET/CT images, although for completeness a section dedicated to alternative heterogeneity metrics can be found in Supplementary material section 1.

Systematically constructing higher-dimensional information from data falls under the general rubric of '-omics', which includes genomics, proteomics and others [19]. Extracting a large number of features from images (including TA metrics, shape descriptors and other quantitative metrics) has become popular under the denomination *radiomics* [20, 21]. The potential of such an approach is to quantify properties of tissues and/or organs beyond the capability of visual interpretation or simple metrics. The use of TA has been widespread in MR and CT imaging since the early 1990s [22, 23] and more recently (end of the 2000s) for PET intratumour heterogeneity characterization [15, 24, 25]. PET images have a priori less-favourable properties for TA than MRI or CT, due to a lower signal-to-noise ratio and spatial resolution, as well as poorer spatial sampling. In addition, reconstructed PET images are often smoothed in clinical practice for visual analysis by clinicians using filters which reduce the textural content of the image, such as the gaussian filter [26]. However, the fidelity and quantitative accuracy of PET imaging has substantially improved in the last decade, with the advent of PET/CT systems, time-of-flight (TOF) capabilities, improved sensitivity, and the incorporation of several quantitative corrections in the current clinical gold standard iterative reconstruction algorithms [27]. Note also that images from the low-dose CT component of PET/CT have different characteristics from higher resolution dosimetry or diagnostic CT images.

In the last 4 years, dozens of studies investigating PET/CT uptake heterogeneity have been published and have mostly focused on the PET component. Unfortunately, especially for TA, the number of required preprocessing steps and the numerous implementation choices in the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

involved workflow, have led to contradictory results and controversies, rendering impossible comparison of results across studies [28]. In addition, the tendency in the biomedical field for positive results to be published more easily than negative ones [29] may have led to over-optimistic interpretation of the potential value of TA in PET. This should also be viewed in the context of current widespread concerns about reproducibility in biomedical research in general [30].

The objective of this review is to provide critical insights into the rapid development of the use of TA to characterize radiotracer uptake heterogeneity in PET/CT images, and identify the issues and challenges that are currently left to address. Finally, some recommendations for future research methodologies and reporting standards are discussed.

The past: clinical potential and underlying motivations

The underlying motivation arose in part from the recognition that standard metrics considered in clinical practice or research studies, i.e. maximum or mean standardized uptake value (SUV_{max} and SUV_{mean}) or the metabolically active tumour volume (MATV), do not fully describe the properties of tumours [14]. Some of these properties, such as shape and uptake heterogeneity, may reflect different tumour profiles associated with their aggressiveness, metastatic potential, or degree of response to a specific treatment, and consequently prognosis [31, 32]. Quantifying these properties could provide indices with higher clinical value than the usual metrics in stratifying patients or identifying poor responders to treatment.

This proof-of-concept regarding the use of TA in the evaluation of PET images was first shown by El Naqa and colleagues in a seminal study in 9 patients with head and neck cancer and 14 patients with cervix cancer [15]. Only two other studies investigating TA in PET were published in the two following years. The first demonstrated the impact of parameters used in PET iterative image reconstruction algorithms on TA metrics, of which many were shown to be sensitive to the resulting varying characteristics of the reconstructed images [24]. This highlighted the need for standardization if such features were to be considered within the context of multicentre trials. Secondly, the extremely high variability (>100 %) observed for some features suggested that they should never be used, even in a single-site, single-scanner study. The second study investigated the predictive value of FDG uptake heterogeneity quantified using TA, in 41 patients with locally advanced oesophageal cancer receiving concomitant chemoradiotherapy, and showed that TA metrics have higher predictive value than SUV [25].

Several subsequent studies have shown significant correlations between the visual assessment of intratumour heterogeneity in PET images by experts and quantitative metrics including the area under the curve of the cumulative histogram [33], shape descriptors [34] (see Supplementary material section 1) and TA metrics [35]. However, the interpretation of the underlying biological meaning of PET image uptake heterogeneity and the explanation of why it may be potentially more powerful than other standard metrics is still largely based on assumptions linking it to differences in underlying metabolism, cellular proliferation, hypoxia and necrosis. This obviously depends on the radiotracer used. However, the vast

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

majority of studies to date have been carried out using FDG (and static SUV images), with only a few examples of the use of other radiotracers, such as FET [36], FLT [5, 37], and DBTZ [16].

By comparison in CT or MRI, there have been several studies linking image-derived TA features with underlying pathophysiological properties including at the level of genomics [7, 11, 38–40], thus providing growing evidence of their relevance and a potential explanation for their observed clinical value. To the best of our knowledge, similar results currently available regarding TA applied to PET are very limited. A study established a correlation between perfusion CT-derived parameters (e.g. blood flow) and TA metrics from FDG PET in stage HI/TV colorectal tumours [41]. Regarding the relationship between PET TA features and data from underlying scales, preliminary results from a prospective study in 54 patients with head and neck cancer have recently been presented, and demonstrate that some PET TA metrics could be linked to altered signalling pathways related, for example, to cell proliferation and apoptosis [42]. Such studies have the potential to help in understanding the observed higher clinical value of these metrics compared to standard quantitative parameters.

The present: an era of rapid expansion

Several dozen studies investigating the clinical value of PET uptake heterogeneity (using TA or other methods) in various tumour types (including oesophageal, lung, rectal, breast, head and neck, and brain cancer, and lymphoma) as well as more recently in neurodegenerative diseases with PET [16, 43] and DAT SPECT [44] have been published in the last 4 years alone. More recently, a few studies have also focused on extracting features from both the PET and CT components (see the section Promising clinical results below). For a more exhaustive list, we refer the reader to other recent reviews [2–4, 9, 45–49]. From a critical review of these studies, several common issues can be identified.

Nomenclature variability, formula and implementation issues

The variability in definitions of TA metrics and nomenclature, as well as errors in methods, published formulae and computational codes complicate any evaluation and comparison of published results. The use of the term “textural” can itself be confusing. In a recent study, the title and abstract refer to “textural parameters”, reporting a higher predictive value regarding response to therapy in a cohort of 27 patients with rectal cancer [50]. However, only first-order histogram-derived features (coefficient of variation, skewness and kurtosis) and none of the textural features of second-order or higher-order features (that actually take into account spatial distribution) were explored. In addition, since these metrics were compared in PET images obtained before, during and after therapy, their repeatability needs careful verification. Yet previous studies have shown a relatively low level of repeatability of these features, especially skewness [51, 52]. This challenge is not restricted to PET studies, as a recent study of the use of DCE MRI in lung cancer made exactly the same use of the term “textural”, although only first-order histogram-derived metrics were used [12]. Another example of potential nomenclature confusion is the use of the term “entropy” to mean “randomness” or “disorder” for the first-order metric, when it is actually the entropy of the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

probability histogram [12]. Furthermore, we recommend using the terms "entropy$_{GLCM}$" and "entropy$_{HIST}$" in order to avoid confusion between the feature calculated in the co-occurrence matrix and that calculated in the histogram, as an intuitive understanding of entropy may not apply to these metrics. Similarly, we recommend using the terms "contrast$_{GLCM}$" and "contrast$_{NGTDM}$" to avoid confusion.

Supplementary material section 3 contains a list of formulae for calculating TA metrics with detailed notes. Several software distributions have also been made available [53, 54], but there is a need to ensure that all feature calculations are accurately implemented before they can be reliably used for research. Supplementary material section 4 contains a list of several such codes with associated remarks.

Workflow complexity

One issue with TA is the very large number of parameters that can theoretically be calculated, in some cases over 100, as well as the number of ways they can be calculated. The recognized sources of variability (acquisition protocol, scanner type, quantitative corrections, type of reconstruction algorithm and parameters, postreconstruction image processing, region of interest definition, etc.) in the standard metrics (SUV, MATV) quantification may also have a similar impact on TA features. There are also additional steps and methodological choices that have a similar (if not higher) impact on the resulting TA metrics. Figure 1 illustrates the complexity of the TA workflow, with the different steps discussed in the following sections. Note than some upstream steps can also have an impact on these choices, such as the segmentation (see section PET tumour volume segmentation below).

First-order features estimate properties of individual voxel values, ignoring the spatial interaction between them (and as such cannot really be considered as "textural" features, because they do not differentiate spatial arrangements and patterns), whereas second-order and higher-order features estimate properties of two or more voxel values occurring at specific locations relative to each other. For these second-order and higher-order TA features, the first steps usually consist of resampling or interpolating the noncubic voxel grids into cubic voxels (seldom carried out) and performing quantization (systematically carried out, also called *discretization, downsampling* or *resampling*) of the original intensities (or SUV) into a discrete set of values. This number determines the size of the matrices that are built and in which TA metrics are subsequently calculated. Several methods have been proposed to perform this quantization (see Supplementary section 2) such as a linear distribution into a set number of bins (e.g. 32 or 64) [15, 25], the use of a clustering algorithm (Max-Lloyd) [55] or into bins of fixed width (e.g. 0.25 SUV [52] or 0.5 SUV [56, 57]). The chosen quantization approach and value can have an important impact on the resulting TA metrics, as well as their relationship with tumour volume or SUV$_{max}$ [51, 56, 58–60], and it is thus an important factor that should not be overlooked, as illustrated in Fig. 2.

The second step consists of building the texture matrices, of which several exist (e.g. grey-level co-occurrence matrix, GLCM; neighbourhood grey tone difference matrix, NGTDM;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and grey level zone size matrix, GLZSM) and which can be built in different ways (see Supplementary material, section 3). For example, co-occurrence matrices quantify relationships between pairs of voxels. They are usually defined according to a given spatial direction and a given distance between the pairs of voxels. For a 3D analysis, 13 directions are often considered and one matrix is built per direction. The TA metric is then calculated in each of these matrices, and the 13 resulting values are averaged. Usually, the distance is set to one voxel. Modifying these choices (e.g. using only one matrix for all directions) can lead to different TA feature distributions (see Fig. 3), associated complementary value with other metrics, and as a consequence, overall clinical value [55, 60].

PET tumour volume segmentation

Numerous studies have used the least robust and/or accurate methods to define overall tumour volume, such as manual delineation or fixed thresholding. Single observer manual delineation suffers from high interobserver and intraobserver variability, whereas fixed thresholding significantly underestimates the true MATV extent by focusing on the tumour subvolume with the highest uptake [61, 62]. This in turn may bias the heterogeneity assessment and the associated ranking of intratumour heterogeneity levels. Another issue concerns the way the tumour volume is a priori considered in the analysis. More specifically one can define functional volume so that areas with low or no radiotracer uptake are included in the volume, or alternatively excluded from it. Excluding these areas would exclude necrotic regions but would also limit the risk of including nonpathological areas in the heterogeneity analysis. The choice of segmentation approach used may result in more or fewer constraints. For example, with a gradient-based tool [63, 64], the resulting contour is binary only and covers the entire tumour including areas without uptake (Fig. 4). On the contrary, with a method based on region growing or clustering paradigms [65, 66], the areas with uptake similar to the background uptake are usually excluded, although they could be included in the analysis with an additional step.

Statistical issues

In the vast majority of published studies no multivariate analysis including potential confounding factors was performed, nor a correction for multiple testing, and very rarely was robust machine learning with cross-validation used. Generally the patient cohorts considered have been very small with respect to the number of explored parameters and tested hypotheses. Ideally, these studies should be combined in a meta-analysis, but because of problems with how results have frequently been reported [21], such a meta-analysis is practically impossible. Finally, the study cohorts have often been heterogeneous in terms of staging or treatment modality, studies have most often been retrospective in nature, and the results have almost never been validated against external cohorts. A recent review of a selection of 15 studies (in both PET and CT) highlighted these issues, and showed that the majority of the studies suffered from at least some of these shortcomings. The review concluded that the clinical value of TA metrics extracted from CT or PET images remains to be demonstrated [28].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Although the bias towards publishing positive results in the biomedical field is strong [29], one has to keep in mind that only a handful of studies have concluded that heterogeneity quantification does not bring any value regarding the clinical endpoint under consideration [67–69], with the overall trend being mostly positive. More specifically, in two studies in cervical cancer a metric based on a "volume versus threshold curve" was not able to predict outcome in 73 patients [67], contradicting a previous assessment in the same cohort [70]. Although other first-order features such as standard deviation, skewness and kurtosis were also included, it should be emphasized that this study essentially highlighted the fact that the metric based on the "volume versus threshold curve" is a surrogate for volume, not a measurement of heterogeneity (see also Supplementary material). The same authors further explored additional metrics (sphericity, extent, Shannon entropy and the accrued deviation from smoothest gradients, i.e. not TA metrics) in another group of 85 patients FIGO stage IIb cancer, with similar negative conclusions regarding the prediction of pelvic lymph node involvement [69]. Finally, in contradiction to the findings of these two studies, the same group also reported the results of another study in which TA metrics had predictive value of response to therapy in 20 patients with cervical cancer when considering their temporal evolution from baseline to week 2, week 4 and post-therapy PET scans [71]. It should also be noted that in all these studies MATV was delineated using a fixed threshold of 40% of $\mathrm{SUV}_{\mathrm{max}}$.

A recent study in breast cancer showed in a prospective homogeneous cohort of 171 women, that contrary to previous results obtained in a smaller cohort ($n = 54$) [72], none of the considered PET TA metrics were able to improve differentiation between the three main molecular subtypes of breast tumours beyond the standard clinical factors and SUV metrics [68].

Redundancy of features

The vast majority of studies were based on analysing a predetermined functional tumour volume, which is thus known prior to the heterogeneity characterization. Therefore a heterogeneity metric can only have complementary (or significantly higher) value if it is not highly correlated with the corresponding volume. The correlation between heterogeneity metrics and the MATV or another image-derived parameter (such as $\mathrm{SUV}_{\mathrm{max}}$, $\mathrm{SUV}_{\mathrm{mean}}$ or total lesion glycolysis = MATV × $\mathrm{SUV}_{\mathrm{mean}}$ [73]) can be explained by two different but complementary aspects: the mathematical/ algorithmic design of the parameter, and the fact that heterogeneity is intrinsically and biologically correlated with MATV. Indeed, the degree of uptake heterogeneity can be expected to be correlated with other tumour properties. In most solid tumours larger volumes exhibit a wider range of heterogeneity patterns and intensity than smaller ones. This is due first to the fact that larger tumours have more potential to be composed of several different types of tissues and regions with variable uptake that can be resolved on PET images than smaller volumes, for which a similar heterogeneity may exist at the cellular and tissue levels but is blurred due to limited spatial resolution. On the other hand, a correlation between high heterogeneity and high SUV seems less logical, since small areas of homogeneous uptake can have high or low $\mathrm{SUV}_{\mathrm{max}}$, whereas both larger homogeneous or heterogeneous lesions can exhibit a wide range of maximum uptake.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The challenge for future studies is therefore to identify which part of the correlation comes from biological reality, from imaging limitations and/or from the mathematical and algorithmic definition of the heterogeneity metrics. Another challenge is to identify the level of correlation above which a TA feature should be excluded from subsequent multivariate analysis since it is unlikely to provide complementary information. This is less trivial than it may sound, since the absolute level of correlation varies depending on the chosen coefficients. Pearson coefficients may significantly underestimate the correlation between two metrics when the correlation is not linear. Kendall and Spearman coefficients both provide rank correlation assessments, but Kendall coefficients are usually smaller than Spearman coefficients, as illustrated in Fig. 5. These would thus require different scales to distinguish strong, moderate and weak correlations (in the case presented in Fig. 5, the correlation is above 0.9 using the Spearman coefficient but is below 0.8 using the Kendall coefficient). Using simple correlation coefficients to select features to combine in a multiparametric model may be suboptimal, and we recommend using robust machine-learning techniques to achieve better redundancy analysis and feature selection/combination.

Although it should be recognized that the relationship between tumour volume and the spatial resolution of PET has an impact on derived metrics, there is an additional factor to take into account for TA metrics that quantify intensity and spatial relationships between pairs or groups of voxels. The correlation between a TA metric and the volume of interest in which it is calculated needs to be analysed not only in terms of absolute volume, but more importantly in terms of the number of voxels involved in the calculation. For example, if we consider a given tumour volume sampled on a 2 × 2 × 2 mm^3 or 4 × 4 × 4 mm^3 grid, the $entropy_{GLCM}$ metric will have a higher value for the 2 × 2 × 2 mm^3 image than for the 4 × 4 × 4 mm^3 image, not because of a higher heterogeneity, but only because of a higher number of voxels involved in the calculation.

This "number of voxels confounding effect" has been demonstrated in a recent study showing that the matrix grid in the reconstruction has a strong impact on most TA metrics [74]. It is especially important to take this into account in a multicentre study where reconstructed matrix sizes vary across sites. Regarding the tumour volume confounding effect, the correlation between TA metrics and tumour volume has been investigated in several studies [59, 60, 75, 76], two of which were specifically focused on the issue. The first found that the minimal volume of interest that would be sufficiently large for the TA to differentiate between different levels of heterogeneity is 45 cm^3 [76]. This study was based on the use of a single TA feature (entropyGLCM) with a uniform quantization ($Q = 152$ bins), an analysis in 2D with two directions (horizontal and vertical), with one co-occurrence matrix used for each direction followed by averaging the resulting values, and without testing different configurations. Another study tackled the same issue by considering 555 tumours of five different types [60]. The correlation between TA metrics and volume was highly variable among TA features, decreased as tumour volumes increased, and depended on the quantization value and design of the co-occurrence matrices. The study also showed that using the same set of parameters as in the previous study [76] led to a very high correlation (>0.9) between the volume and $entropy_{GLCM}$, whereas using a different configuration (a smaller quantization value of 64 and only one cooccurrence matrix taking

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

into consideration all directions simultaneously instead of averaging the values obtained in different matrices) the correlation dropped below 0.5. With this calculation, entropy$_{GLCM}$ thus provided complementary information to volume (Fig. 3), which also translated into complementary prognostic value: combining the two parameters improved the stratification of 101 patients with non-small-cell lung cancer (NSCLC) in terms of overall survival [60]. These two studies clearly highlight the need to carefully consider interactions between tumour volume and heterogeneity quantitative metrics. The latter study also showed that the choices made in the calculation workflow can determine the efficacy of a metric.

Beyond the relationships with tumour volume, TA provides the ability to calculate numerous parameters that also exhibit high levels of redundancy [59, 75]. It is therefore necessary to establish a method of selection amongst all calculated features (and amongst all ways of calculating them). The properties of an “ideal” heterogeneity metric as well as the recommended methodology to assess them are listed in Table 1.

Repeatability and reproducibility/robustness

Repeatability is a measure of precision under identical or near-identical conditions, e.g. double-baseline (also called test–retest) studies. Reproducibility, in contrast, is a measure of precision when location, measuring system or other factors differ [79]. In reproducibility studies, the objective is to measure the effects of different conditions on the performance of a quantitative imaging biomarker with the goal of demonstrating equivalent performance under less-restrictive study conditions. Repeatability has been studied for TA in PET, and it has been found that only a handful of metrics have repeatability limits similar to MATV and SUV measurements, and are therefore reliable enough to be considered further, especially in the context of therapy monitoring [51, 52, 57]. Regarding reproducibility or robustness, it has been shown that only a few features are robust with regard to variations in the type of image reconstruction algorithm [24, 52, 74, 80]. It has also shown that the effects of tumour segmentation [52, 57, 58, 75], postreconstruction smoothing [26], quantization [52, 58] and partial volume effect correction [75] vary among TA metrics.

More recently, the effects of respiratory motion have been investigated in two studies by comparing TA features on PET images of lung cancer patients with and without respiratory gating, and the studies showed that TA features may be affected in standard nongated acquisitions, particularly in the lower lung lobes [81, 82]. It has also been shown that the use of a respiration-averaged CT scan instead of a helical CT scan for attenuation correction of the PET data has a greater effect on SUV and total lesion glycolysis than on TA metrics [83]. Another study demonstrated that TA features calculated in parametric maps derived from dynamic PET acquisitions or from corresponding static SUV images are not significantly different, suggesting that heterogeneity quantification on parametric images using TA may not provide significant additional information compared to that provided by static SUV images [84]. Finally, it has also been shown that even basic stochastic effects of PET acquisitions can affect some TA metrics [85]. All these factors may affect not only the absolute values of calculated features, but also their correlation with volume. Figure 6 illustrates this in a set of tumours reconstructed using two different voxel sizes (the only modified parameter in the reconstruction). Using either nearest-neighbour or B-spline

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

interpolation of the image with larger voxels exactly the same number of voxels was obtained as in the image with smaller voxels.

Promising clinical results

Demonstration of the clinical value of TA requires large cohorts of patients and rigorous statistical analysis. Although a number of studies have included only between 20 and 70 patients [50, 71, 86–92], some of the most recent studies have included between 80 and more than 200 patients: 88 patients with oropharyngeal squamous cell carcinoma [93], 103 with bone and soft tissue lesions [94], 101 with early-stage NSCLC [95], 112 with oesophageal cancer and 101 with NSCLC [60], 113 with glioma [36], 107 and 217 with oesophageal cancer [96, 97], 132 with lymph node involvement in lung cancer [98], 116, 195 and 201 with NSCLC [99–101], 137 with pancreatic lesions [102], and 188 lesions in lymphoma patients [103]. Some of the most recent studies have also used more robust statistical analysis, compared to these recently reviewed [28], several of them using a machine-learning method, e.g. neural networks [96], support vector machines [94, 98, 103] or the least absolute shrinkage and selection operator (LASSO) [95, 101]. The majority of these recent studies have concluded that TA can provide useful quantitative metrics regarding patient management (prognosis, response to therapy, distant metastasis prediction) in different cancer models except one that showed more mixed results [104], whereas another concluded that the improvement, although significant, may not be sufficient to have a clinical impact [97].

A few recent studies have investigated the potential combination of image-derived features from both PET and the corresponding low-dose CT component of the PET/CT dataset [8, 94, 98, 99, 103, 105], while in another recent study TA from PET and MRI were combined to predict lung metastases in soft-tissue sarcomas of the extremities [55]. A recent proof-of-concept study (in two patients) has investigated the characterization of renal cell carcinoma in simultaneous ^{18}F-FLT PET/MRI acquisitions (with images obtained before and during treatment) [106]. Finally, the combination of PET image-derived features and other contextual data may be considered. For example, in a recent study the combination of zone-size nonuniformity extracted from pretreatment FDG PET and key immunohistochemistry metrics led to improved stratification in 113 patients with advanced stage oropharyngeal squamous cell carcinoma [107]. In another recent study, the combination of TA metrics extracted from both PET and low-dose CT components of PET/CT standard acquisitions led to higher prognostic stratification power than clinical staging [99].

When dealing with a large number of variables and cohorts of limited size, robust methods for feature selection combined with an appropriate classifier and testing with cross-validation can provide tools with good performance. However, validation using an external cohort remains the gold standard, although it is still rarely performed in PET/CT studies. In a recent study, 31 of 101 patients were used for validation, and the first 70 for building the model [95]. Contrary to the field of PET/CT, the use of machine-learning techniques including robust feature selection, the combination of features within classifiers and cross-validation or validation in an external cohort are well established in the fields of CT [7, 77, 108, 109] and MR [110–114] radiomics.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To date, only a limited number of clinical studies investigating TA in PET/CT have exploited up-to-date techniques from the field of machine learning. Recently, a more technical study showed that appropriate hierarchical forward selection of features combined with a support vector machine classifier can improve results [78]. A recent study focused on this topic and compared 14 feature selection methods and 12 classifiers using large CT datasets, ranking relative performance in terms of the accuracy and stability of each approach [77]. Interestingly, the performance was mostly affected by the choice of classifier (34 % of total variance). Although not yet performed on PET datasets, this work could nonetheless form the basis for selecting appropriate machine-learning methods for future studies investigating the value of TA in PET imaging.

Conclusions: a future for texture analysis in PET?

Although not impossible, it is challenging to compare the results from the numerous studies currently available and draw concrete conclusions as to the clinical value of TA in PET imaging. This is due to large variability in the implemented methodology associated with the workflow complexity involved in the calculation of features, in combination with the lack of technical details provided in most studies. In addition, numerous issues related to the statistical analysis and bias in publishing mostly positive results further increase the difficulty in drawing firm conclusions from the currently published literature. Keeping in mind these current limitations, it should nonetheless be emphasized that most currently published studies indicate that extracting more advanced image features from medical images, including PET/CT, provides complementary and additional value. Although the level of evidence is probably still insufficient, a positive trend can be observed. Thus the use of TA in PET/CT images should not be abandoned but rather reinforced by increasing the required level of conduct of the studies to enable the field to move forward. We suggest the following several objectives that we should consider as a community to achieve this:

1. Organize and develop a benchmark standard for TA metrics. This would include standardized physical and digital reference objects, open-source verified formulae and codes tested again reference objects, and expected values/results for comparison.

2. Generate and circulate draft recommendations on image preprocessing, analysis and standardization, especially within the context of multicentre studies. Convene consensus groups to review, revise and ratify.

3. Establish recommendations on methodological choices regarding the calculation of TA metrics and identify repeatable, reproducible and meaningful features (as well as their optimal calculation).

4. Share publicly available cohorts of patients with PET/CT images and associated clinical data, along with clinical endpoints (survival, response to therapy, tumour type classification, etc.) so that research groups can test/evaluate their workflow.

5. Support larger prospective multicentre studies and the use of robust statistical analysis by exploiting the methods from the field of machine learning.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

6. Adopt standards for publishing methods and results such as those promoted by the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network [115].

7. Advocate for improved peer review, insisting on at least one "statistical reviewer" with knowledge of machine-learning methodologies.

Establishing a benchmark (objective 1) could start by providing users with tools to validate feature calculation codes. Test images with known and verified associated metrics values for a range of calculation methods and choices could be provided, so that users can validate their TA metric implementations [85].

The next step (objective 2) should consist of generating and circulating draft recommendations and guidelines regarding choices in preprocessing and segmentation steps, especially within the context of multicentre studies. Although variability inherent in merging results based on images from various devices and reconstructions algorithms from different vendors may be difficult to avoid, some recommendations can already be made based on current results. First, preprocess and resample images to a common voxel size, preferably isotropic [55]. This would allow major issues in comparing co-occurrence-derived metrics calculated with different spatial sampling to be avoided [55, 74]. Second, avoid postreconstruction smoothing altogether, or use appropriate edge-preserving filters [26]. Third, automated segmentation approaches with robustness against heterogeneous distributions should be used [61, 66]. Alternatively, if only fixed or adaptive thresholding methods are available, manual/visual checking and editing should be mandatory to avoid under-segmentation of areas of heterogeneous uptake.

Objective 3 could be achieved by establishing recommendations as to which features should preferably be used and which excluded (e.g. features that have been identified as having very poor repeatability and robustness), and recommendations for workflow choices to obtain features with the lowest redundancy and highest clinical value. In this regard, we provide in the Supplementary material a list of verified formulae for the usual features, as well as comments, corrections and implementation recommendations to avoid common mistakes and misconceptions. We also recommend that preferably features that have been demonstrated as robust and repeatable be relied upon [51, 52, 57, 58, 75].

Finally, beyond providing test images, objects, and open-source codes and formulae to improve standardization between research groups (objective 1), it would be beneficial to make available a benchmark containing publicly available clinical datasets of PET/CT images along with clinical end-point information (prognosis, response to therapy, tumour type, etc.) and other clinical data (objective 4). This would allow any research group to test its own workflow (image preprocessing, tumour segmentation, TA metric calculation, machine-learning feature selection and classifiers, etc.) and then compare its results with those of other groups. The Cancer Imaging Archive (TCIA) could support such efforts, as it already contains several publicly available cohorts of patients with images from various modalities and the associated clinical data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Further efforts along these lines could be organized within task groups of the EANM, QIBA, QIN, SNMMI and AAPM and this should be pursued as soon as possible if TA (and *radiomics* in general) are to have any future in PET/CT imaging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Issam El Naqa, Philippe Lambin, Hugo Aerts, Ralph Leijenaar, Floris Van Velden, Martin Vallières, Arman Rahmim, Matt Nyflot and Art Chaovalitwongse, as well as the members of the RSNA Quantitative Imaging Biomarkers Alliance and the NCI Quantitative Imaging Network for many helpful discussions.

Funding Supported in part by NIH grant U01-CA148131.

This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the “Investing for the Future” program under reference ANR-10-LABX-07-01. With the support of the National Institute of Cancer (INCa project #C14020NS).

References

1. Gerlinger M, Rowan AT, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892. [PubMed: 22397650]
2. 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; 3:573–589. [PubMed: 23093486]
3. Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJ. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. Eur J Nucl Med Mol Imaging. 2013; 40:133–140. [PubMed: 23064544]
4. 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; 21:249–257. [PubMed: 25421725]
5. Willaime JM, Turkheimer FE, Kenny LM, Aboagye EO. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. Phys Med Biol. 2013; 58:187–203. [PubMed: 23257054]
6. Weber WA, Schwaiger M, Avril N. Quantitative assessment of tumor metabolism using FDG-PET imaging. Nucl Med Biol. 2000; 27:683–687. [PubMed: 11091112]
7. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5:4006. [PubMed: 24892406]
8. Win T, Miles KA, Janes SM, Ganeshan B, Shastry M, Endozo R, et al. Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non-small cell lung cancer. Clin Cancer Res. 2013; 19:3591–3599. [PubMed: 23659970]
9. Asselin MC, O’Connor JP, Boellaard R, Thacker NA, Jackson A. Quantifying heterogeneity in human tumours using MRI and PET. Eur J Cancer. 2012; 48:447–455. [PubMed: 22265426]
10. 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; 105:139–145. [PubMed: 21673686]
11. Nicolasjilwan M, Hu Y, Yan C, Meerzaman D, Holder CA, Gutman D, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. J Neuroradiol. 2015; 42:212–221. [PubMed: 24997477]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

12. Yoon SH, Park CM, Park SJ, Yoon J-H, Hahn S, Goo JM. Tumor heterogeneity in lung cancer: assessment with dynamic contrast-enhanced MR imaging. Radiology. 2016

13. Michallek F, Dewey M. Fractal analysis in radiological and nuclear medicine perfusion imaging: a systematic review. Eur Radiol. 2014; 24:60–69. [PubMed: 23974703]

14. O'Sullivan F, Roy S, Eary J. A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. Biostatistics. 2003; 4:433–448. [PubMed: 12925510]

15. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit. 2009; 42:1162–1171. [PubMed: 20161266]

16. Gonzalez ME, Dinelle K, Vafai N, Heffernan N, McKenzie J, Appel-Cresswell S, et al. Novel spatial analysis method for PET images using 3D moment invariants: applications to Parkinson's disease. Neuroimage. 2013; 68:11–21. [PubMed: 23246861]

17. van Velden FH, Cheebsumon P, Yaqub M, Smit EF, Hoekstra OS, Lammertsma AA, et al. Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. Eur J Nucl Med Mol Imaging. 2011; 38:1636–1647. [PubMed: 21617975]

18. Ganeshan B, Miles KA. Quantifying tumour heterogeneity with CT. Cancer Imaging. 2013; 13:140–149. [PubMed: 23545171]

19. Micheel, CM.Nass, SJ., Omenn, GS., editors. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, Institute of Medicine. Evolution of translational omics: lessons learned and the path forward. Washington (DC): National Academies Press (US); 2012. http://www.ncbi.nlm.nih.gov/books/NBK202168/

20. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012; 48:441–446. [PubMed: 22257792]

21. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016; 278:563–577. [PubMed: 26579733]

22. Mir AH, Hanmandlu M, Tandon SN. Texture analysis of CT-images for early detection of liver malignancy. Biomed Sci Instrum. 1995; 31:213–217. [PubMed: 7654965]

23. Schad LR, Blüml S, Zuna I. MR tissue characterization of intracranial tumors by means of texture analysis. Magn Reson Imaging. 1993; 11:889–896. [PubMed: 8371644]

24. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010; 49:1012–1016. [PubMed: 20831489]

25. 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; 52:369–378. [PubMed: 21321270]

26. Hatt M, Hanzouli H, Rest CCL, Visvikis D. Comparison of edge-preserving filters for unbiased quantification in 18F-FDG PET imaging. J Nucl Med. 2015; 56:1828. [PubMed: 26429956]

27. Vaquero JJ, Kinahan P. Positron emission tomography: current challenges and opportunities for technological advances in clinical and preclinical imaging systems. Annu Rev Biomed Eng. 2015; 17:385–414. [PubMed: 26643024]

28. 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:e0124165. [PubMed: 25938522]

29. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. PLoS One. 2013; 8:e66844. [PubMed: 23861749]

30. Ioannidis JPA. How to make more published research true. PLoS Med. 2014; 11:e1001747. [PubMed: 25334033]

31. Basu S, Kwee TC, Gatenby R, Saboury B, Torigian DA, Alavi A. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. Eur J Nucl Med Mol Imaging. 2011; 38:987–991. [PubMed: 21451997]

32. Visvikis D, Hatt M, Tixier F, Rest CLC. The age of reason for FDG PET image-derived indices. Eur J Nucl Med Mol Imaging. 2012; 39:1670–1672. [PubMed: 22968400]

33. Watabe T, Tatsumi M, Watabe H, Isohashi K, Kato H, Yanagawa M, et al. Intratumoral heterogeneity of F-18 FDG uptake differentiates between gastrointestinal stromal tumors and abdominal malignant lymphomas on PET/CT. Ann Nucl Med. 2012; 26:222–227. [PubMed: 22187313]

34. Kim DH, Jung JH, Son SH, Kim CY, Jeong SY, Lee SW, et al. Quantification of intratumoral metabolic macroheterogeneity on 18F-FDG PET/CT and its prognostic significance in pathologic N0 squamous cell lung carcinoma. Clin Nucl Med. 2016; 41:e70–e75. [PubMed: 26284762]

35. Tixier F, Hatt M, Valla C, Fleury V, Lamour C, Ezzouhri S, et al. Visual versus quantitative assessment of intratumor 18F-FDG PET uptake heterogeneity: prognostic value in non-small cell lung cancer. J Nucl Med. 2014; 55:1235–1241. [PubMed: 24904113]

36. Pyka T, Gempt J, Hiob D, Ringel F, Schlegel J, Bette S, et al. Textural analysis of pre-therapeutic [18F]-FET-PET and its correlation with tumor grade and patient survival in high-grade gliomas. Eur J Nucl Med Mol Imaging. 2016; 43:133–141. [PubMed: 26219871]

37. Majdoub, M., Visvikis, D., Tixier, F., Hoeben, B., Visser, E., Cheze Le Rest, C., et al. Proliferative 18F–FLT PET tumor volumes characterization for prediction of locoregional recurrence and disease-free survival in head and neck Cancer; Presented at the Society of Nuclear Medicine and Molecular Imaging Annual Meeting; 8–12 June 2013; Vancouver, Canada.

38. 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; 25:675–680. [PubMed: 17515910]

39. Gevaert O, Mitchell LA, Achrol AS, Xu J, Echegaray S, Steinberg GK, et al. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. Radiology. 2014; 273:168–174. [PubMed: 24827998]

40. Wan T, Bloch BN, Plecha D, Thompson CL, Gilmore H, Jaffe C, et al. A radio-genomics approach for identifying high risk estrogen receptor-positive breast cancers on DCE-MRI: preliminary results in predicting OncotypeDX risk scores. Sci Rep. 2016; 6:21394. [PubMed: 26887643]

41. Tixier F, Groves AM, Goh V, Hatt M, Ingrand P, Le Rest CC, et al. Correlation of intra-tumor 18F-FDG uptake heterogeneity indices with perfusion CT derived parameters in colorectal cancer. PLoS One. 2014; 9:e99567. [PubMed: 24926986]

42. Tixier F, Hatt M, Rest CCL, Simon B, Key S, Corcos L, et al. Signaling pathways alteration involved in head and neck cancer can be identified through textural features analysis in 18F-FDG PET images: a prospective study. J Nucl Med. 2015; 56:449.

43. Klyuzhin IS, Gonzalez M, Shahinfard E, Vafai N, Sossi V. Exploring the use of shape and texture descriptors of positron emission tomography tracer distribution in imaging studies of neurodegenerative disease. J Cereb Blood Flow Metab. 2015

44. Rahmim A, Salimpour Y, Jain S, Blinder SA, Klyuzhin IS, Smith GS, et al. Application of texture analysis to DAT SPECT imaging: relationship to clinical assessments. Neuroimage Clin. 2016

45. Hatt M, Tixier F, Rest CLC, Visvikis D. Nouveaux indices en TEP/TDM: mythe et réalités. Med Nucl. 2015; 39:331–338.

46. Carlier T, Bailly C. State-of-the-art and recent advances in quantification for therapeutic follow-up in oncology using PET. Front Med. 2015; 2:18.

47. Houshmand S, Salavati A, Hess S, Werner TJ, Alavi A, Zaidi H. An update on novel quantitative techniques in the context of evolving whole-body PET imaging. PET Clin. 2015; 10:45–58. [PubMed: 25455879]

48. Rahim MK, Kim SE, So H, Kim HJ, Cheon GJ, Lee ES, et al. Recent trends in PET image interpretations using volumetric and texture-based quantification methods in nuclear oncology. Nucl Med Mol Imaging. 2014; 48:1–15. [PubMed: 24900133]

49. Cheng NM, Fang YH, Yen TC. The promise and limits of PET texture analysis. Ann Nucl Med. 2013; 27:867–869. [PubMed: 23943197]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

50. Bundschuh RA, Dinges J, Neumann L, Seyfried M, Zsótér N, Papp L, et al. Textural parameters of tumor heterogeneity in 18F-FDG PET/CT for therapy response assessment and prognosis in patients with locally advanced rectal cancer. J Nucl Med. 2014; 55:891–897. [PubMed: 24752672]
51. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med. 2012; 53:693–700. [PubMed: 22454484]
52. van Velden FHP, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol. 2016
53. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015; 42:1341–1353. [PubMed: 25735289]
54. Fang YH, Lin CY, Shih MJ, Wang HM, Ho TY, Liao CT, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. BiomedRes Int. 2014; 2014:248505.
55. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015; 60:5471–5496. [PubMed: 26119045]
56. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015; 5:11075. [PubMed: 26242464]
57. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013; 52:1391–1397. [PubMed: 24047337]
58. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. Eur Radiol. 2015; 25:2805–2812. [PubMed: 25994189]
59. Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med. 2014; 55:414–422. [PubMed: 24549286]
60. Hatt M, Majdoub M, Vallières 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; 56:38–44. [PubMed: 25500829]
61. Hatt M, Cheze-le Rest C, van Baardwijk A, Lambin P, Pradier O, Visvikis D. Impact of tumor size and tracer uptake heterogeneity in (18)F-FDG PET and CT non-small cell lung cancer tumor delineation. J Nucl Med. 2011; 52:1690–1697. [PubMed: 21990577]
62. Dong X, Wu P, Sun X, Li W, Wan H, Yu J, et al. Intra-tumour 18F-FDG uptake heterogeneity decreases the reliability on target volume definition with positron emission tomography/ computed tomography imaging. J Med Imaging Radiat Oncol. 2015; 59:338–345. [PubMed: 25708154]
63. Geets X, Lee JA, Bol A, Lonneux M, Gregoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging. 2007; 34:1427–1438. [PubMed: 17431616]
64. Nelson A, Brockway K, Nelson A, Piper J. PET tumor segmentation: validation of a gradient-based method using a NSCLC PET phantom. J Nucl Med. 2009; 50(Suppl 2):1659.
65. Hofheinz F, Langner J, Petr J, Beuthien-Baumann B, Steinbach J, Kotzerke J, et al. An automatic method for accurate volume delineation of heterogeneous tumors in PET. Med Phys. 2013; 40:082503. [PubMed: 23927348]
66. Hatt M, Rest ClC, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys. 2010; 77:301–308. [PubMed: 20116934]
67. Brooks FJ, Grigsby PW. Current measures of metabolic heterogeneity within cervical cancer do not predict disease outcome. Radiat Oncol. 2011; 6:69. [PubMed: 21658258]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

68. Groheux D, Majdoub M, Tixier F, Le Rest CC, Martineau A, Merlet P, et al. Do clinical, histological or immunohistochemical primary tumour characteristics translate into different (18)F-FDG PET/CT volumetric and heterogeneity features in stage II/III breast cancer? Eur J Nucl Med Mol Imaging. 2015; 42:1682–1691. [PubMed: 26140849]

69. Brooks FJ, Grigsby PW. FDG uptake heterogeneity in FIGO IIb cervical carcinoma does not predict pelvic lymph node involvement. Radiat Oncol. 2013; 8:294. [PubMed: 24365202]

70. Kidd EA, Grigsby PW. Intratumoral metabolic heterogeneity of cervical cancer. Clin Cancer Res. 2008; 14:5236–5241. [PubMed: 18698042]

71. Yang F, Thomas MA, Dehdashti F, Grigsby PW. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. Eur J Nucl Med Mol Imaging. 2013; 40:716–727. [PubMed: 23340594]

72. Soussan M, Orlhac F, Boubaya M, Zelek L, Ziol M, Eder V, et al. Relationship between tumor heterogeneity measured on FDG-PET/CT and pathological prognostic factors in invasive breast cancer. PLoS One. 2014; 9:e94017. [PubMed: 24722644]

73. Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging The visual response score and the change in total lesion glycolysis. Clin Positron Imaging. 1999; 2:159–171. [PubMed: 14516540]

74. Yan J, Lim JC-S, 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; 56:1667–1673. [PubMed: 26229145]

75. Hatt M, Tixier F, Rest CLC, Pradier O, Visvikis D. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur J Nucl Med Mol Imaging. 2013; 40:1662–1671. [PubMed: 23857457]

76. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med. 2014; 55:37–42. [PubMed: 24263086]

77. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015; 5:13087. [PubMed: 26278466]

78. Mi H, Petitjean C, Dubray B, Vera P, Ruan S. Robust feature selection to predict tumor treatment outcome. Artif Intell Med. 2015; 64:195–204. [PubMed: 26303106]

79. Sullivan DC, Obuchowski NA, Kessler LG, Raunig DL, Gatsonis C, Huang EP, et al. Metrology standards for quantitative imaging biomarkers. Radiology. 2015; 277:813–825. [PubMed: 26267831]

80. Quantitative Imaging Biomarkers Alliance, FDG-PET/CT Technical Committee. FDG-PET/CT as an imaging biomarker measuring response to cancer therapy, version 1.05. RSNA. 2013

81. Yip S, McCall K, Aristophanous M, Chen AB, Aerts HJ, Berbeco R. Comparison of texture features derived from static and respiratory-gated PET images in non-small cell lung cancer. PLoS One. 2014; 9:e115510. [PubMed: 25517987]

82. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer. Transl Oncol. 2015; 8:524–534. [PubMed: 26692535]

83. Cheng NM, Fang YH, Tsan DL, Hsu CH, Yen TC. Respiration-averaged CT for attenuation correction of PET images – impact on PET texture features in non-small cell lung cancer patients. PLoS One. 2016; 11:e0150509. [PubMed: 26930211]

84. Tixier F, Vriens D, Cheze-Le Rest C, Hatt M, Disselhorst JA, Oyen WJ, et al. Comparison of tumor uptake heterogeneity characterization between static and parametric 18F-FDG PET images in non-small cell lung cancer. J Nucl Med. 2016

85. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. J Med Imaging (Bellingham). 2015; 2:041002. [PubMed: 26251842]

86. Pyka T, Bundschuh RA, Andratschke N, Mayer B, Specht HM, Papp L, et al. Textural features in pre-treatment [F18]-FDG-PET/ CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. Radiat Oncol. 2015; 10:100. [PubMed: 25900186]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

87. Cook GJ, O'Brien ME, Siddique M, Chicklore S, Loi HY, Sharma B, et al. Non-small cell lung cancer treated with erlotinib: heterogeneity of (18)F-FDG uptake at PET-association with treatment response and prognosis. Radiology. 2015; 276:883–893. [PubMed: 25897473]

88. Mu W, Chen Z, Liang Y, Shen W, Yang F, Dai R, et al. Staging of cervical cancer based on tumor heterogeneity characterized by texture features on (18)F-FDG PET images. Phys Med Biol. 2015; 60:5123–5139. [PubMed: 26083460]

89. Oh JS, Kang BC, Roh JL, Kim JS, Cho KJ, Lee SW, et al. Intratumor textural heterogeneity on pretreatment (18)F-FDG PET images predicts response and survival after chemoradiotherapy for hypopharyngeal cancer. Ann Surg Oncol. 2015; 22:2746–2754. [PubMed: 25487968]

90. Cheng NM, Fang YH, Chang JT, Huang CG, Tsan DL, Ng SH, et al. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. J Nucl Med. 2013; 54:1703–1709. [PubMed: 24042030]

91. Lovinfosse P, Janvary ZL, Coucke P, Jodogne S, Bernard C, Hatt M, et al. FDG PET/CT texture analysis for predicting the outcome of lung cancer treated by stereotactic body radiation therapy. Eur J Nucl Med Mol Imaging. 2016

92. Yip SS, Coroller TP, Sanford NN, Mamon H, Aerts HJ, Berbeco RI. Relationship between the temporal changes in positronemission-tomography-imaging-based textural features and pathologic response and survival in esophageal cancer patients. Front Oncol. 2016; 6:72. [PubMed: 27066454]

93. Cheng N-M, Fang Y-HD, Lee L, Chang JT-C, Tsan D-L, Ng S-H, et al. Zone-size nonuniformity of 18F-FDG PET regional textural features predicts survival in patients with oropharyngeal cancer. Eur J Nucl Med Mol Imaging. 2015; 42:419–428. [PubMed: 25339524]

94. Xu R, Kido S, Suga K, Hirano Y, Tachibana R, Muramatsu K, et al. Texture analysis on (18)F-FDG PET/CT images to differentiate malignant and benign bone and soft-tissue lesions. Ann Nucl Med. 2014; 28:926–935. [PubMed: 25107363]

95. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW, et al. Early-stage non-small cell lung cancer: quantitative imaging characteristics of (18)F Fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology. 2016

96. 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:e0137036. [PubMed: 26355298]

97. van Rossum PS, Fried DV, Zhang L, Hofstetter WL, van Vulpen M, Meijer GJ, et al. The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. J Nucl Med. 2016; 57:691–700. [PubMed: 26795288]

98. Gao X, Chu C, Li Y, Lu P, Wang W, Liu W, et al. The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from (18)F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. Eur J Radiol. 2015; 84:312–317. [PubMed: 25487819]

99. Desseroit MC, Visvikis D, Tixier F, Majdoub M, Guillevin R, Perdrisot R, et al. Development of a nomogram combining clinical staging with 18F-FDG PET/CT image features in non-small-cell lung cancer stage I-III. Eur J Nucl Med Mol Imaging. 2016

100. Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III non-small cell lung cancer: prognostic value of FDG PET quantitative imaging features combined with clinical prognostic factors. Radiology. 2016; 278:214–222. [PubMed: 26176655]

101. Ohri N, Duan F, Snyder BS, Wei B, Machtay M, Alavi A, et al. Pretreatment 18FDG-PET textural features in locally advanced non-small cell lung cancer: secondary analysis of ACRIN 6668/RTOG 0235. J Nucl Med. 2016

102. Hyun SH, Kim HS, Choi SH, Choi DW, Lee JK, Lee KH, et al. Intratumoral heterogeneity of 18F-FDG uptake predicts survival in patients with pancreatic ductal adenocarcinoma. Eur J Nucl Med Mol Imaging. 2016

103. Lartizien C, Rogez M, Niaf E, Ricard F. Computer-aided staging of lymphoma patients with FDG PET/CT imaging based on textural information. IEEE J Biomed Health Inform. 2014; 18:946–955. [PubMed: 24081876]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

104. Bang JI, Ha S, Kang SB, Lee KW, Lee HS, Kim JS, et al. Prediction of neoadjuvant radiation chemotherapy response and survival using pretreatment [(18)F]FDG PET/CT scans in locally advanced rectal cancer. Eur J Nucl Med Mol Imaging. 2016; 43:422–431. [PubMed: 26338180]

105. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiother Oncol. 2012; 102:239–245. [PubMed: 22098794]

106. Antunes J, Viswanath S, Rusu M, Valls L, Hoimes C, Avril N, et al. Radiomics analysis on FLT-PET/MRI for characterization of early treatment response in renal cell carcinoma: a proof-of-concept study. Transl Oncol. 2016; 9:155–162. [PubMed: 27084432]

107. Wang HM, Cheng NM, Lee LY, Fang YH, Chang JT, Tsan DL, et al. Heterogeneity of (18) F-FDG PET combined with expression of EGFR may improve the prognostic stratification of advanced oropharyngeal carcinoma. Int J Cancer. 2016; 138:731–738. [PubMed: 26311121]

108. 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. [PubMed: 26697407]

109. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim JH, Choi YL, et al. Decoding tumor phenotypes for ALK, ROS1, and RET fusions in lung adenocarcinoma using a radiomics approach. Medicine (Baltimore). 2015; 94:e1753. [PubMed: 26469915]

110. Upadhaya T, Morvan Y, Stindel E, Le Reste PJ, Hatt M. A framework for multimodal imaging-based prognostic model building: preliminary study on multimodal MRI in glioblastoma multiforme. IRBM. 2015; 36:345–350.

111. Wang J, Kato F, Oyama-Manabe N, Li R, Cui Y, Tha KK, et al. Identifying triple-negative breast cancer using background parenchymal enhancement heterogeneity on dynamic contrast-enhanced MRI: a pilot radiomics study. PLoS One. 2015; 10:e0143308. [PubMed: 26600392]

112. Cameron A, Khalvati F, Haider M, Wong A. MAPS:a quantitative radiomics approach for prostate cancer detection. IEEE Trans Biomed Eng. 2016; 63:1145–1156. [PubMed: 26441442]

113. Khalvati F, Wong A, Haider MA. Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. BMC Med Imaging. 2015; 15:27. [PubMed: 26242589]

114. Sharma RR, Marikkannu P. Hybrid RGSA and support vector machine framework for three-dimensional magnetic resonance brain tumor classification. ScientificWorldJournal. 2015; 2015:184350. [PubMed: 26509188]

115. Pandis N, Fedorowicz Z. The international EQUATOR network: enhancing the quality and transparency of health care research. J Appl Oral Sci. 2011

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

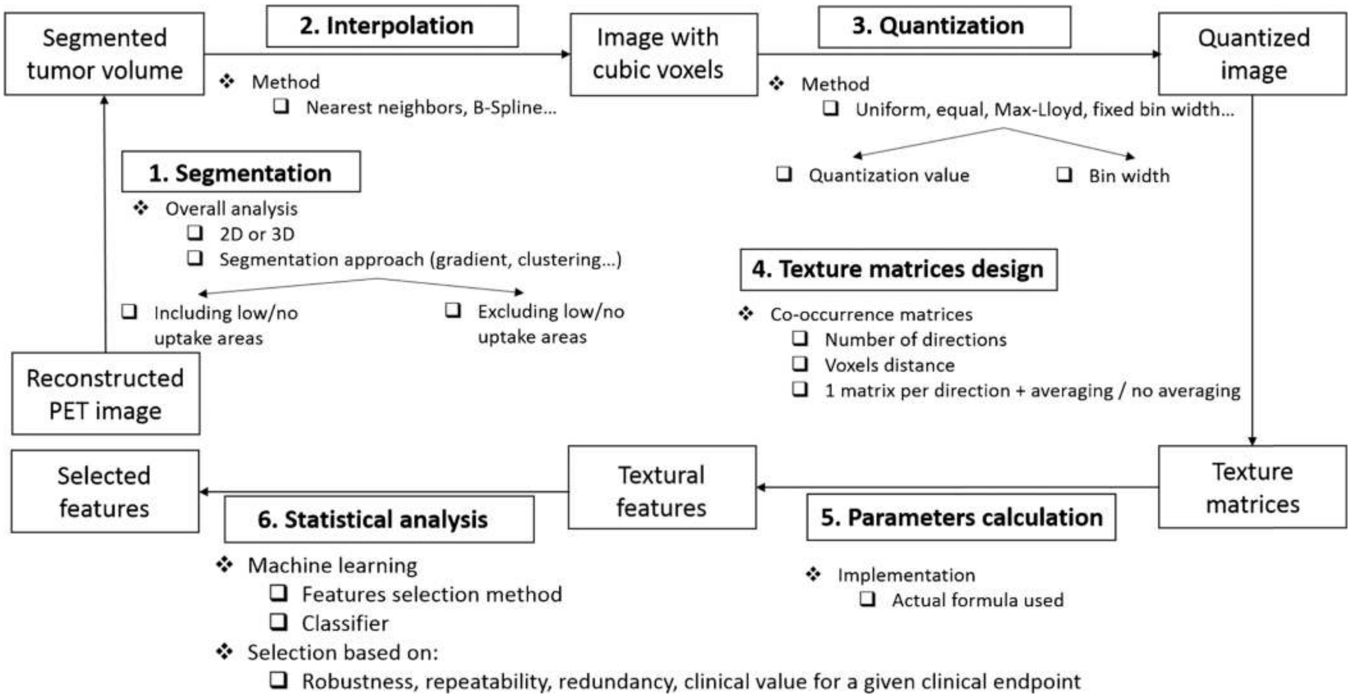

Fig. 1.

Workflow involved in the calculation and selection of texture analysis from a reconstructed PET image

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

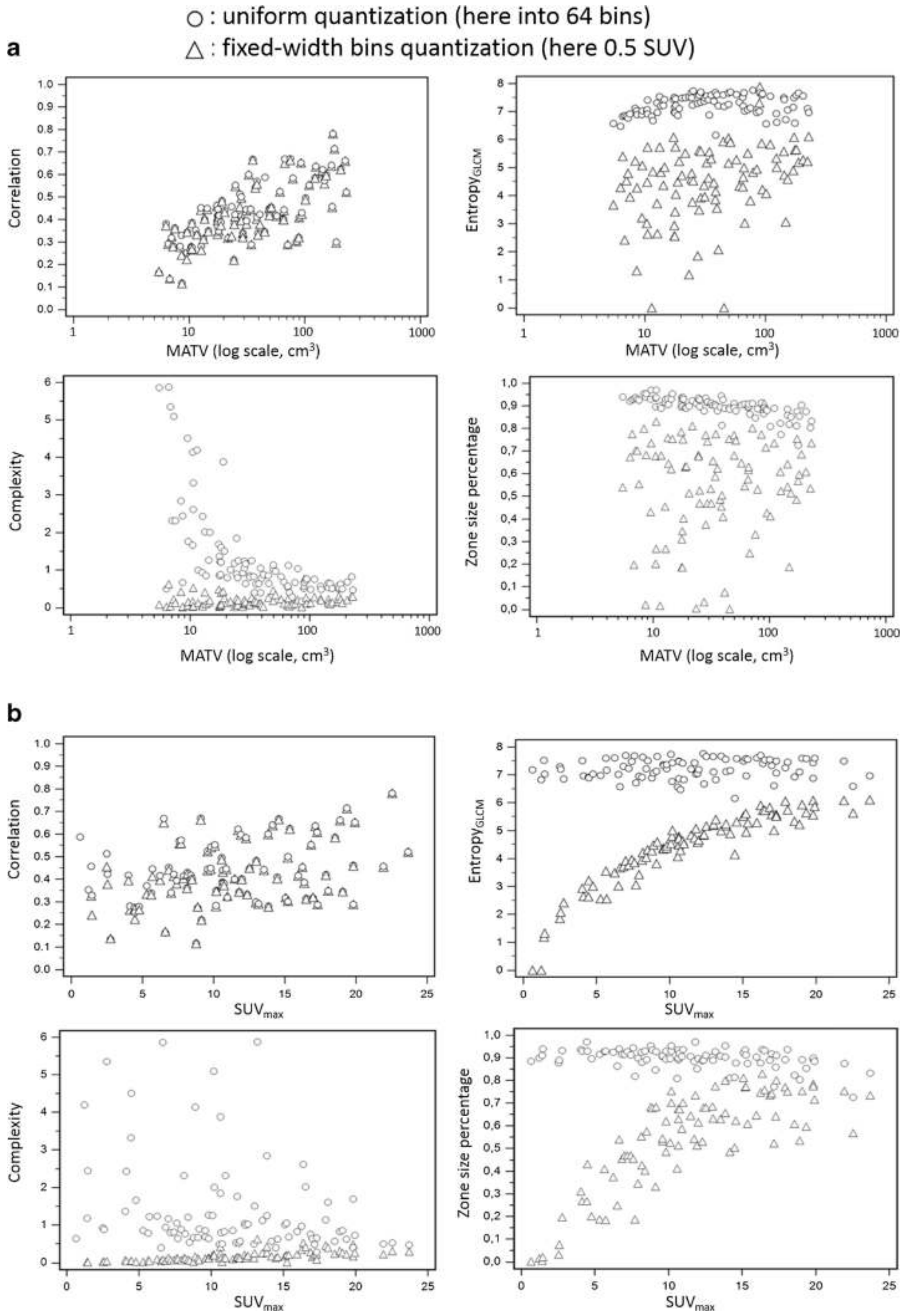

Fig. 2.

Distributions with respect to (**a**) MATV and (**b**) SUV_{max} of four ► TA features (correlation and $entropy_{GLCM}$ from GLCM, complexity from NGTDM and zone size percentage from GLZSM) calculated after either quantization into a set number of bins (here 64) or into bins of fixed width (here 0.5 SUV). Note that correlation is not affected, compared to the three other metrics. Also, note the inverted correlation with MATV and SUV_{max}, when changing the quantization approach

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

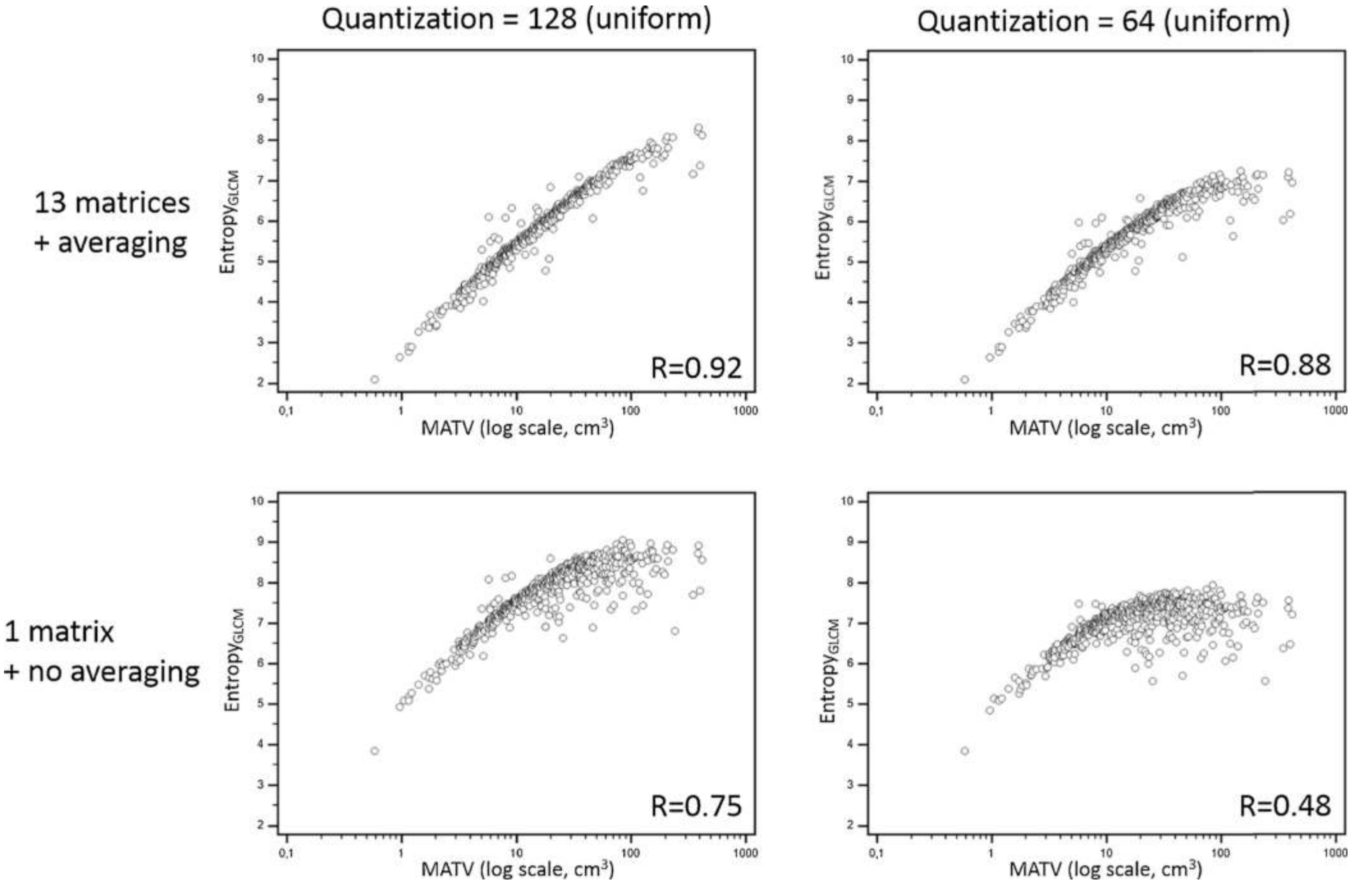

Fig. 3.

Distribution of heterogeneity as measured with entropy$_{GLCM}$ with respect to MATV for 555 lesions in five tumour types, according to four different configurations: with quantization of either 64 or 128 grey levels (uniformly distributed) and using either one single co-occurrence matrix (without averaging) or 13 matrices followed by averaging

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

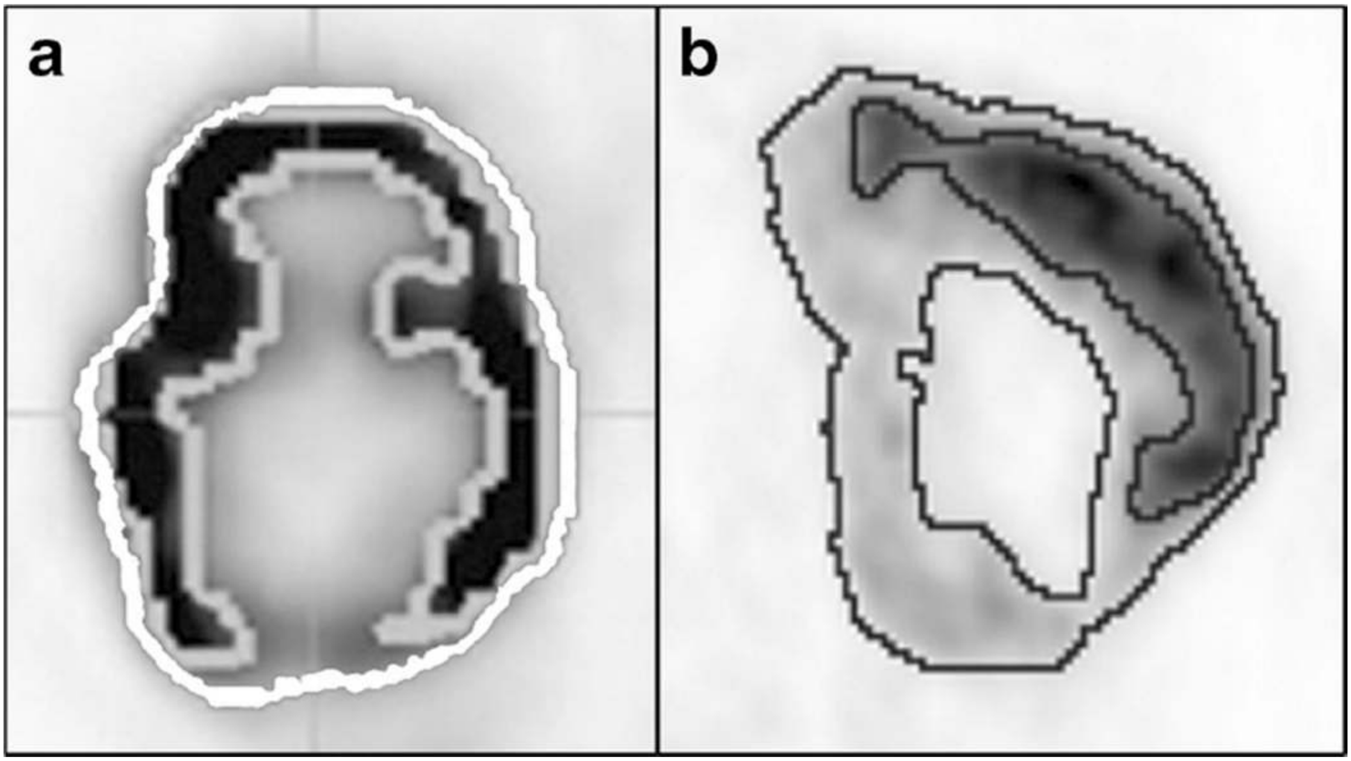

Fig. 4.

Trade-offs in segmentation results using (**a**) a contour-based approach (PETedge from MIMVista software, *white external contour*) or (**b**) a clustering-based approach (the FLAB algorithm, *black countours*). In **a** the *light grey contour* inside the tumour corresponds to a fixed threshold. In **b** the various areas with different uptake levels are automatically determined and may be included or excluded from the heterogeneity analysis at the cost of higher complexity

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

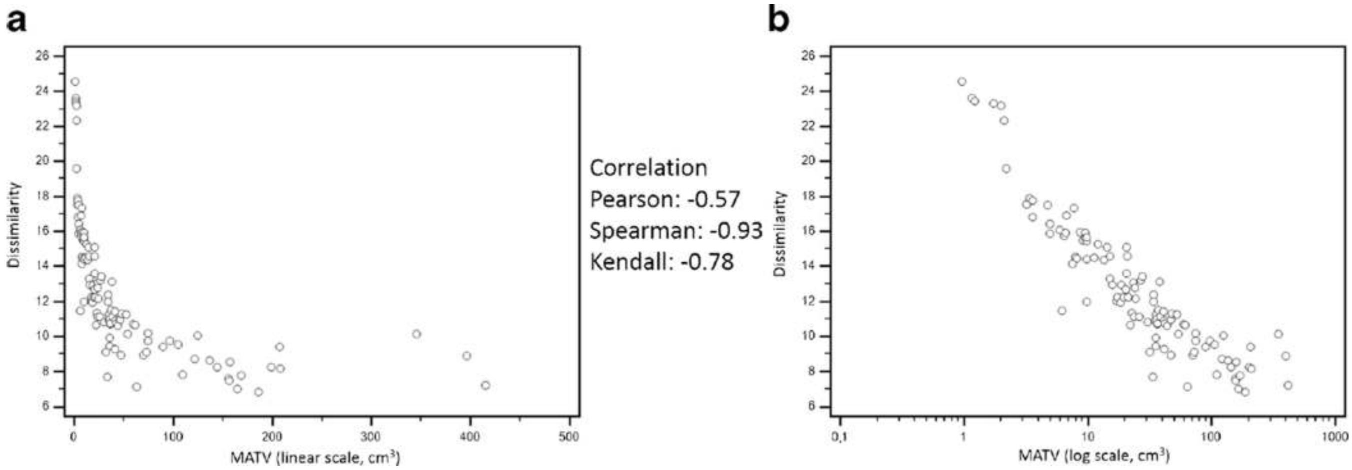

Fig. 5.

Relationship between a textural feature and the corresponding MATV in 116 patients with non-small-cell lung cancer (**a** linear scale, **b** log scale) and the resulting quantification of the correlation according to Pearson coefficients and different rank coefficients (Spearman and Kendall)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

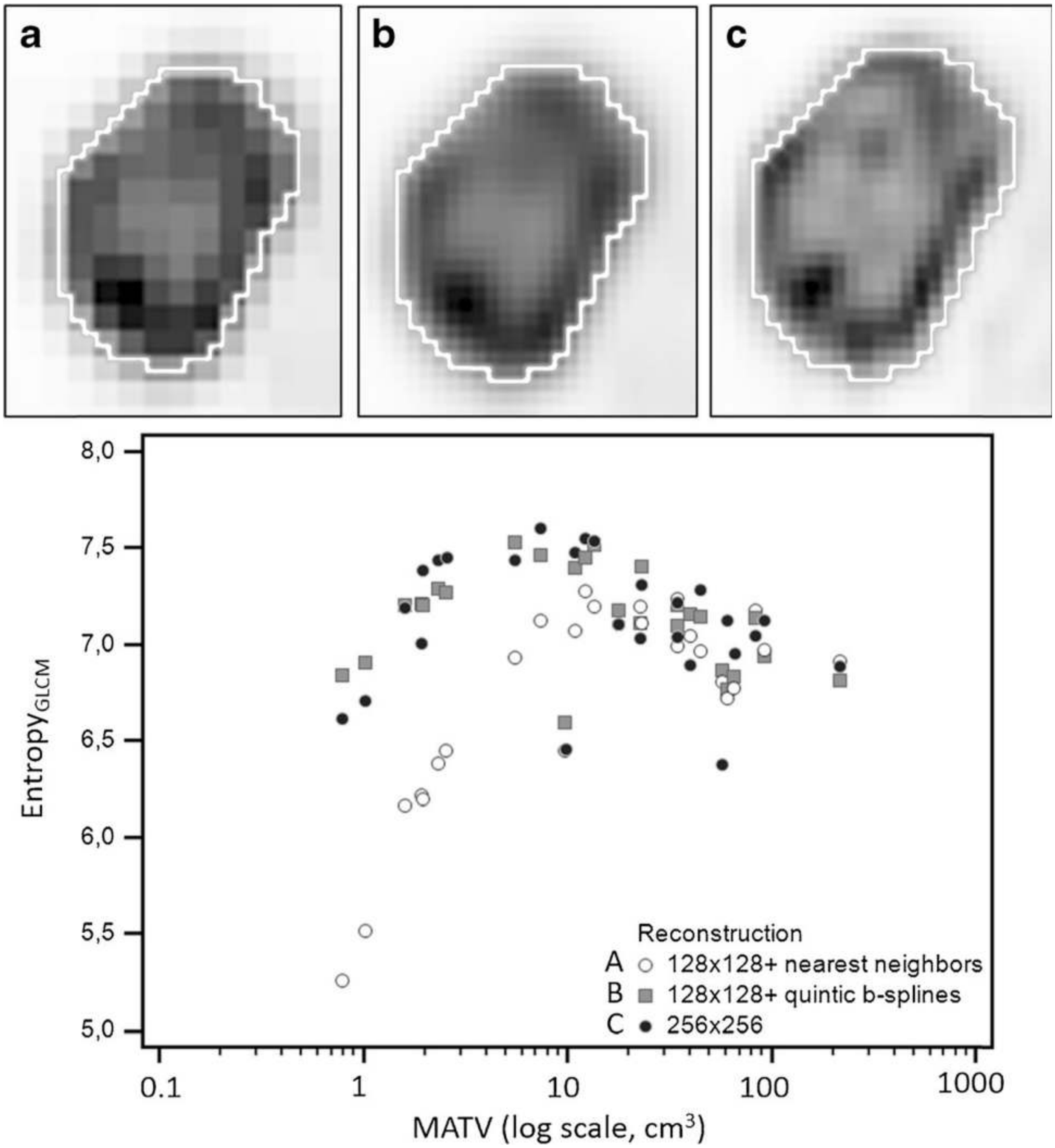

Fig. 6.

Distributions of heterogeneity (entropy$_{GLCM}$) with respect to MATV for a set of 25 tumours reconstructed with either 4 × 4 × 4 mm^3 voxels (**a, b**) or 2 × 2 × 2 mm^3 voxels (c). The image with 4 × 4 × 4 mm^3 voxels was upsampled to 2 × 2 × 2 mm^3 voxels using either nearest neighbour interpolation (**a**) or B-spline interpolation (**b**)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Properties of an ideal heterogeneity metric

Property	Recommended methodology for evaluation
Repeatable	Compare metrics calculated on test-retest PET/CT images [51, 57] using, for example, the Bland-Airman method
Reproducible/robust	Compare metrics calculated through various analysis pipelines (with/without preprocessing such as denoising or partial volume effect correction, various segmentation approaches…) [57, 75]
Least redundant with other TA metrics (and other variables)	Quantify and rank statistical correlations between features [59, 60, 75]. Use machine-learning techniques to select features and combine them with other variables [77, 78]
Offers value in regard to a given clinical endpoint	Quantify correlation with response to treatment, diagnosis, survival, differentiation of tumour types using robust machine-learning techniques for classification, logistic regression and multivariate analysis, and learning/testing in separate cohorts (at least considering leave-one-out cross-validation) [28, 77, 78]