Quantification of Heterogeneity as a Biomarker in Tumor Imaging: A Systematic Review

Lejla Alic1,2*, Wiro J. Niessen1,3, Jifke F. Veenland1

1 Biomedical Imaging Group Rotterdam, Department of Radiology and Medical Informatics, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands, 2 Department of Intelligent Imaging, Netherlands Organization for Applied Scientific Research (TNO), The Hague, The Netherlands, 3 Imaging Physics, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands

Abstract

Background: Many techniques are proposed for the quantification of tumor heterogeneity as an imaging biomarker for differentiation between tumor types, tumor grading, response monitoring and outcome prediction. However, in clinical practice these methods are barely used. This study evaluates the reported performance of the described methods and identifies barriers to their implementation in clinical practice.

Methodology: The Ovid, Embase, and Cochrane Central databases were searched up to 20 September 2013. Heterogeneity analysis methods were classified into four categories, i.e., non-spatial methods (NSM), spatial grey level methods (SGLM), fractal analysis (FA) methods, and filters and transforms (F&T). The performance of the different methods was compared.

Principal Findings: Of the 7351 potentially relevant publications, 209 were included. Of these studies, 58% reported the use of NSM, 49% SGLM, 10% FA, and 28% F&T. Differentiation between tumor types, tumor grading and/or outcome prediction was the goal in 87% of the studies. Overall, the reported area under the curve (AUC) ranged from 0.5 to 1 (median 0.87). No relation was found between the performance and the quantification methods used, or between the performance and the imaging modality. A negative correlation was found between the tumor-feature ratio and the AUC, which is presumably caused by overfitting in small datasets. Cross-validation was reported in 63% of the classification studies. Retrospective analyses were conducted in 57% of the studies without a clear description.

Conclusions: In a research setting, heterogeneity quantification methods can differentiate between tumor types, grade tumors, and predict outcome and monitor treatment effects. To translate these methods to clinical practice, more prospective studies are required that use external datasets for validation: these datasets should be made available to the community to facilitate the development of new and improved methods.

Introduction

Tumors are often inhomogeneous. Regional variations in cell death, metabolic activity, proliferation and vascular structure are observed. There is increasing evidence that solid tumors may consist of subpopulations of cells with different genotypes and phenotypes [1]. These distinct populations of cancer cells can interact in a competitive way [2] and may differ in sensitivity to treatments [3,4]. This heterogeneity can be detected using diagnostic imaging techniques at a genetic, molecular or cellular level [4,5], or at a cell population level. The advantage of diagnostic imaging techniques is their non-invasive nature and the fact that the whole tumor is taken into account, whereas cellular diagnostic techniques are invasive and limited to a discrete set of tumor samples. Various imaging techniques are available to visualize the heterogeneity in tissue characteristics, such as necrosis, metabolic activity, cell density and vascularization. Observed heterogeneity in an image is a reflection of the phenotypic variation of the tumor and is reported to be associated with underlying gene-expression patterns [6].

Image heterogeneity can be quantified using a variety of texture analysis methods. As such, image heterogeneity is potential biomarker for tumor characterization, for response prediction and monitoring. Parameters in hot spots, as quantified with dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), are more relevant for monitoring tumor response than parameters averaged over the whole tumor [7–9]. When a region of the tumor is not well vascularized or is hypoxic, chemotherapy and radiotherapy are more likely to fail. The existence of poorly vascularized or hypoxic areas within a tumor is an important component of tumor radiation resistance and correlates with treatment failure [10]. In radiotherapy, the heterogeneity can be used to guide treatment [11,12]: an ongoing trial is currently escalating the dose to the part of the tumor with high standardized uptake values [13]. Also for computed tomography (CT), image heterogeneity has prognostic value [6].
Several methods are available to quantify tumor heterogeneity from imaging data. Many studies have used histogram-derived features such as percentile values, standard deviation (SD) and enhancing fraction. However, these features do not take into account the spatial distribution of the intensity values. In contrast, texture methods take spatial information into account by quantifying the spatial variations in the images. Ideally, these methods are independent of the absolute signal intensities in the image. They provide additional and independent information (such as the average signal intensity) compared to histogram-derived measures. These methods result in features which can be considered to be imaging biomarkers providing information on the underlying tumor heterogeneity. Some of these features are related to image properties that are visually perceived by the radiologist, whereas others are more abstract [14].

By means of a systematic review, the aim of this study is to investigate the performance of different heterogeneity imaging biomarkers extracted from diagnostic tumor images for differentiation between tumor types, tumor grading, outcome prediction and treatment monitoring.

The following research questions were formulated:

- Which analysis methods are used to quantify heterogeneity or texture in tumor imaging, with the aim to differentiate between tumor types, tumor grading, outcome prediction and treatment monitoring?
- What are the reported performances of the different analysis methods? Is there a relation between performance and analysis method?

What is the potential clinical impact of the methods? Can the performance results be generalized? Is the performance evaluated in addition to established imaging biomarkers?

## Methods

### Data Sources and Search method

This review was performed in accordance with the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) guidelines [13], with details summarized in Checklist S1. In January 2013 the study protocol was registered with the International Prospective Register of Systematic Reviews (Identification number: CRD4201300634) [16]. A systematic search was conducted in the databases of Medline, Embase, and Cochrane Central. The search was performed with the aid of an experienced librarian on September 20th 2013.

The following topics were used for the searches:

1. Neoplasms
2. Heterogeneity, texture
3. MRI, MRS, CT, PET, SPECT, ultrasonography
4. Differentiation between tumor types, tumor grading, classification, staging, treatment response, survival, and treatment outcome

Full details of the Embase search is included in Text S1. The results from all three searches were combined and verified to ensure exclusion of publications containing the same title, written by the same authors, and published in the same journal. The remaining publications were considered for study selection.

### Study Selection

Two authors (L.A. and J.F.V.) independently reviewed the titles and abstracts. The selected publications then underwent full-text screening. During the title and abstract review, any discrepancies about study inclusion were resolved by full-text screening. Any discrepancies during the following stages were resolved by discussion. The bibliographies of seminal review papers [17–19] were reviewed to identify additional relevant articles.

### Inclusion and exclusion criteria

We included only publications related to diagnostic imaging which reported quantification of tumor heterogeneity or tumor texture with the goal to differentiate between tumor types, tumor grading, outcome prediction and tumor response monitoring. No restrictions were made based on location, type, stage or grade of malignancy. Prior to review, a decision was made to exclude any study with too few participants, i.e., for patient studies (n<10) and for animal studies (n<5). Therefore, all case studies, and studies with no information on the number of subjects, were excluded. In addition, all the following types of studies were excluded:

- publications based on non-tumor images;
- publications not based on quantitative assessment of heterogeneity or texture in images
- publications without one of the following goals: differentiation between tumor types, tumor grading, or outcome prediction or treatment monitoring;
- publications not based on in vivo studies (histology, phantom, ex vivo, synthetic data);
- publications describing non-original research (editorial, letter to the editor, review, meta-analysis, opinion publications).

### Data extraction

A data extraction form was designed. All selected publications were independently reviewed and data extraction was cross-checked. Disagreements between the reviewers were resolved by consensus. The following data were extracted from the full papers: year of publication, human or animal study, type of study (retrospective or prospective), number of subjects, number of tumors, location of tumor, imaging modality, tracer/contrast agent, goal of heterogeneity/texture analysis, and type of heterogeneity/texture quantification method used. For studies reporting on the same analysis method based on the identical dataset, only the latest publication was included. For publications reporting classification experiments, the following data were extracted: number of candidate heterogeneity features, dimensionally reduction technique used, number of selected features used in the best classification experiment, the results of the best classification experiment, i.e., accuracy, sensitivity, specificity, area under the receiver operator curve (AUC), type of cross-validation used, and use of an external validation set. For publications using statistical hypothesis testing the following data were extracted: the number of candidate features, and the number of features that showed a significant difference between outcome categories (before and after Holm-Bonferroni correction) [20]. All publications were divided into two categories:

- Publications reporting cross-sectional measurements with the aim to differentiate between tumor types, tumor grading, and treatment outcome prediction.
- Publications reporting longitudinal measurements for tumor treatment monitoring.
Data synthesis and analysis

The imaging modalities were summarized into four categories: i) magnetic resonance imaging (MRI), ii) computed tomography (CT), iii) positron emission tomography (PET), single photon emission computed tomography (SPECT), and iv) ultrasonography (US). No further subdivision was made regarding the type of imaging protocol or use of contrast agent.

Image analysis methods to estimate tumor heterogeneity were divided into four categories: non-spatial methods, local spatial distribution methods, fractal analysis, and a category consisting of filters and transforms.

Non-spatial methods (NSM). These methods characterize tumor heterogeneity by non-spatial descriptors, such as descriptors of the gray-level frequency distributions: standard deviation, skewness, maximum, minimum, range, peak height, peak position, and percentile values.

Spatial gray-level methods (SGLM). Methods included in the second category extract the local spatial image intensity distribution. This category includes gray-tone spatial-dependence matrix (GTSDM) [21], neighborhood gray-tone difference matrix (NGTDM) [22], run-length matrix (RLM), and Local Binary Pattern (LBP) [23]. The GTSDM, originally proposed by Haralick et al. [21], is often referred to as co-occurrence or the second-order histogram. When divided by the total number of neighboring pixels in the image, this matrix becomes the estimate of the joint probability of two pixels at a distance along a given direction having a particular gray value. The NGTDM, originally proposed by Amadas and King [22], is based on spatial changes in gray values by inspecting the difference between gray levels of a specific pixel and the average gray level of their surrounding neighbors. The RLM, originally proposed by Galloway [24], is subsidiary to the observation that a coarse texture would have relatively longer gray level runs compared to a fine texture. This matrix provides information about runs of pixels with the same gray level values in a given direction. LBP, originally proposed by Ojala et al. [25] and later modified to a rotation and scale invariant approach [23], represents local texture. In its simplest form it labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number.

Fractal analysis (FA). The third category consists of FA methods that overcome the scale problem by providing a statistical measure reflecting pattern changes as a function scale. The two basic parameters in FA are fractal dimension (FD) and lacunarity [26]. An often used method to estimate FD is box counting [26]. This procedure systematically overlays an image with a series of grids with increasing/decreasing size. For each step this procedure captures the predefined relevant features [27]. Another frequently used technique in FA is the blanket method [26], which is often used in its extended form, as described by Peleg et al. [28]. This method estimates the surface area by measuring the volume between an upper and lower blanket.

Filters and Transforms (F&T). The fourth category consists of a collection of image processing algorithms that extract texture features. Examples are methods that use techniques defined in the spatial domain such as filters (Gabor filters or Law’s filters) or transformations to other domains (Fourier transform, Wavelet transform, S-transform, discrete cosine transform). Since the various methods have only been used in a limited number of publications included in the present review, these methods were grouped together.

Publications reporting classification experiments. Publications were considered classification studies if they reported a classification result such as accuracy, sensitivity, specificity or AUC values. Only publications in which the results of the classification experiments were solely based on texture parameters were further analyzed. These studies often utilize a high number of candidate features to describe a tumor. When the number of extracted features is too large to perform a statistically meaningful classification [29], the extracted features can be redundant in the information they retain. Because an increase of dimensionality in the feature space results in an increase of its volume, the feature space is sparsely filled. The use of an extensive number of features for classification purposes can result in over-fitting, which reduces the possibility of generalization; this paradox is generally referred to as the ‘curse of dimensionality’ [30].

To keep the system manageable, dimensionality reduction techniques were commonly applied to select a subset of features that were relevant for the classification problem. The ratio between the number of tumors classified and the dimensionality of the feature space (e.g., the number of selected features) should be chosen in a meaningful way. In pattern recognition applications, the rule of thumb is to use 5–10 datasets per feature per category [31]. Therefore, we evaluated the number of candidate features, the number of selected features, and the ratio between the number of tumors included in the study and the number of selected classification features. A one-way ANOVA was used to test for differences in classification results between the modalities and analysis methods.

Publications reporting on significance testing. A commonly used approach to test the validity of the selected features is significance testing. For heterogeneity analysis, many publications compute a large number of features. As multiple comparisons generally require a stronger level of evidence to be considered significant, the Holm-Bonferroni correction [20] can be applied. This correction allows for the significance levels for single and multiple comparisons to be directly comparable. In these publications, we evaluated whether a Holm-Bonferroni correction was applied and, if this was not the case, computed the number of significant features after correction using the available data. A one-way ANOVA was used to test for differences in the number of significant features, before and after Holm-Bonferroni correction, between the modalities and the analysis methods used.

Results

Figure 1 presents details on the literature search. In summary, of the 7351 potentially relevant articles, 480 (6.5%) were considered for inclusion after abstract review. After these latter papers had undergone full-text screening, an additional 249 publications were excluded. The remaining 231 original publications entered the data extraction phase. In this phase an additional 22 papers [32–53] were excluded as they reported results of a similar analysis method on the same dataset as that used in another paper; for these publications, the most recent one was included in the analysis. Finally, data from 209 studies [7,14,54–228] were extracted for further analysis.

General characteristics

Table 1 presents the characteristics of the included publications (after removing duplicate publications). A publication may include more than one imaging modality, analysis method, or goal. Two studies (1%) reported on two imaging modalities, and 66 studies (32%) reported on two or more analysis methods.

Since 2008, the number of imaging studies quantifying tumor heterogeneity has been steadily increasing, i.e. from 8 papers in 2006–2007 to 66 publications in 2012–2013 (figure 2-A). Prior to 2006, heterogeneity was mainly studied based on US data.
Since 2007, most studies quantifying tumor heterogeneity are based on MRI. Generally, the non-spatial method (NSM) and the spatial gray-level method (SGLM) are the most frequently used to analyze tumor heterogeneity (Figure 2-C). Although the number of publications using these methods has increased since 2007, their contribution to heterogeneity literature is relatively stable. The number of studies reporting tumor response monitoring has varied over the years, ranging from 0–20% (Figure 2-D).

Breast tumors were studied in 33% (n = 69) of the publications. Figure 3 shows the distribution of studies per tumor location. Figure 3-A shows the use of imaging modalities for quantification of tumor heterogeneity per primary tumor location. MRI is used primarily for brain and breast tumors, CT for lung and...
gastrointestinal tumors, PET for gastrointestinal, lung tumors and sarcoma, and US for breast tumors. Heterogeneity analysis of brain tumors was performed almost exclusively with MRI, while for breast tumors both MRI and US were used.

Figure 3-B presents the analysis methods used per primary tumor location. For almost all locations, all methods were used. For prostate, breast, and head and neck analysis, the SGLM was the most frequently used. For all other locations, the NSM was the favored modality. Heterogeneity analyses for longitudinal studies were mainly performed for gastrointestinal and breast tumors (Figure 3-C).

Figure S1 summarizes the publications included in the present review (n = 209) in a matrix form. The publications are divided into different imaging modalities and analysis methods, and are available for download for each cell separately. Each cell in the matrix links to the supplementary EndNote file containing the records for these publications.

Figure 4-A shows the relation between imaging modality and analysis methods for cross-sectional studies. In general, 74% of these studies used either MRI or US. The SGLM (37%) and NSM (36%) are most frequently used to grade and diagnose tumors. Figure 4-B shows the relation between imaging modality and analysis method for the longitudinal studies (n = 27). MRI was

---

**Table 1.** Characteristics of the included publications (n = 209).

| Characteristic            | n   | %    |
|---------------------------|-----|------|
| Imaging method            |     |      |
| MRI                       | 75  | 36%  |
| CT                        | 40  | 19%  |
| PET                       | 14  | 7%   |
| US                        | 81  | 39%  |
| Analysis method           |     |      |
| NSM                       | 121 | 58%  |
| SGLM                      | 103 | 49%  |
| FA                        | 21  | 10%  |
| F&T                       | 58  | 28%  |
| Study goal                |     |      |
| Diagnosis/grading/outcome pred. | 182 | 87%  |
| Response monitoring       | 27  | 13%  |
| Study type                |     |      |
| Retrospective             | 118 | 56%  |
| Retrospective (with inclusion criteria) | 63 | 30%  |
| Prospective               | 28  | 13%  |
| Type of subjects          |     |      |
| Human                     | 197 | 94%  |
| Animal                    | 12  | 6%   |
| Type of experiment        |     |      |
| Classification            | 139 | 67%  |
| Significance testing      | 64  | 30%  |
| Neither                   | 6   | 3%   |

Imaging modalities: magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), ultrasonography (US).
Analysis methods: non-spatial methods (NSM), spatial grey level methods (SGLM), fractal analysis (FA) methods, and filters and transforms (F&T).
doi:10.1371/journal.pone.0110300.t001

Figure 2. Number of publications reporting on tumor heterogeneity analysis for all publications bi-annually. Total number of publications (A), publications per imaging modality (B), publications per analysis method (C), and publications per goal (D).
doi:10.1371/journal.pone.0110300.g002
Figure 3. Publications reporting on quantification of tumor heterogeneity in cancer sites summarized for imaging modality (A), analysis method (B), and study aim (C). Publications can report on more than one analysis method. The acronyms used: Gyn – gynecological, H&N – head and neck, GIST – gastrointestinal. doi:10.1371/journal.pone.0110300.g003
used in 70% of these studies and PET in 11%. In 7% of the studies, US-based heterogeneity quantification was used for tumor response monitoring. NSM is the most frequently used (69%) analysis method in longitudinal studies.

A relatively small number of all studies (13%) utilized a prospective study design. Figure 5-A shows the relation between imaging modality and analysis method used for cross-sectional studies (n = 12). US is the most frequently used modality, whereas NSM is the most frequently used analysis method. Figure 5-B shows the relation between imaging modality and analysis method for publications reporting longitudinal studies (n = 16). Again, most data were analyzed with NSM. In contrast to MRI, CT, US and PET are rarely used for heterogeneity quantification in prospective longitudinal studies.

Publications reporting classification experiments

Of all included studies, 67% (n = 139) reported classification experiments and 30% reported significance testing. The remaining 3% either did not report quantitative results or the experiments were not completely described. Also, 23 studies only reported results of classification experiments where the texture features were combined with non-texture features. For these latter publications, it was not possible to extract the performance of the texture features separately and, therefore, these results were excluded from further analysis. Additionally, 10 studies were excluded because the number of generated or selected features was lacking. Of the papers reporting classification experiments (n = 106), 45% used US, 37% used MRI, 13% used CT, and 5% used PET. In 42% of the classification papers, features originating from different texture analysis methods were combined. Some studies reporting classification experiments (n = 39) performed no feature reduction, and the median number of candidate features used in these studies (6) was significantly lower than that of candidate features in the studies using feature reduction techniques (38). The remaining 67 studies reporting classification experiments used one of the methods commonly applied in statistics, pattern recognition, or machine learning. These methods were summarized into three categories: filters, wrappers and embedded methods.

Figure 6 shows the relation between the number of candidate features and the number of selected features used in classification experiments for different imaging modalities (Figure 6-A) and different analysis methods (Figure 6-B). For the papers presented on the dotted line, no feature selection was performed. The number of candidate features ranged from 1–5280 (median 22) while the number of selected features ranged from 1–476 (median 3). The distribution of the numbers of selected features can be assessed as boxplots for imaging modality (Figure 6-C) and for analysis methods (Figure 6-D).

About 63% of the publications describing a classification experiment, reported cross-validation or training test sets as a technique to limit the effect of over-fitting on the available data. Figure 6-B shows that the combination of features from different methods generally leads to a higher number of candidate features. In general, in publications reporting the use of more than one analysis method more extensive feature reduction is applied compared to publications reporting on the use of the separate analysis methods.

In the classification experiments, one or more of the following performance measures were reported: sensitivity, specificity, accuracy, or AUC. Figure 7-A shows the AUC per imaging modality and Figure 7-B the AUC per analysis method. The differences in performance (as measured by AUC) are shown in Figure 7-C per imaging modality and in Figure 7-D per analysis method.

The supplementary material provides the figures for accuracy (Figure S2), sensitivity (Figure S3) and specificity (Figure S4) per imaging modality and per analysis method. In these figures, the reported performance is depicted as a function of the tumor-feature ratio (ratio between the number of tumors included and the number of selected features). In general, the tumor-feature ratio...
ratio ranged from 0.46–502 (median 20) with (on average) 29% of the publications showing a tumor-feature ratio ≤10.

With respect to the analysis method, publications using the F&T, or a combination of methods, had the highest risk of a tumor-feature ratio ≤10, i.e. 53% and 42%, respectively. With regard to imaging modality, CT publications had the highest percentage (43%) with a tumor-feature ratio <10.

Using a one-way ANOVA, no significant differences were found in the performance measures between the modalities or between the analysis methods used. However, there was a negative correlation between the logarithm of the number of tumors per selected feature and the AUC (r = −0.32, p<0.05) and the specificity (r = −0.48, p<0.05).

Publications using statistical hypothesis testing

Of all included studies, 30% (n = 64) reported statistical hypothesis testing with the number of features ranging from 1–320 (median 4). Of these studies, 39% were based on MRI, 26%, on CT, 14% on PET, and 21% on US. Similarly, in 61% of the cases, data were analyzed using NSM, 12% using SGLM, 3% using FA, 6% using F&T, and 18% using a combination of these methods. The number of significant features, as reported by the authors, ranged from 0–76 (median 1). Since multiple comparisons generally require a stronger level of evidence to be considered significant, the Holm-Bonferroni correction [20] was applied by the original research authors, or by the authors of this review paper. This correction allows direct comparison to be made of the significance levels of single and multiple comparisons. For eight papers the correction could not be performed due to missing information. After the Holm-Bonferroni correction, the number of significant features ranged from 0–6 (median 1). Figure 8 shows the number of significant features before and after the Holm-Bonferroni correction per imaging modality (A) and per analysis method (B). In 45% of the papers the number of significant features decreased after correction. Using a one-way ANOVA, no significant differences were found in the number of significant features between the modalities. With respect to the analysis method used, a one-way ANOVA established a significant difference in the number of significant features (p<0.018). Publications using SGLM reported more significant features. However, after the Holm-Bonferroni correction, the numbers of significant features were similar between all analysis methods used.

Figure 6. Number of features used in classification experiments for different imaging modalities (A) and for different analysis methods (B). Boxplot representing distribution in number selected features for imaging modality (C) and for analysis methods (D). To enhance visibility, we excluded for both boxplots two studies with large numbers of selected features.
doi:10.1371/journal.pone.0110300.g006
Figure 7. The AUC for different imaging modalities (A) and for different analysis methods (B) as a function of tumor-feature ratio in the classification experiments. The scatter plot shows each imaging modality and analysis method separately. Dotted line represents the ratio of 10 tumors per selected feature. Boxplot representing distribution in AUC for imaging modality (C) and for analysis methods (D).

doi:10.1371/journal.pone.0110300.g007

Figure 8. Number of significant features before and after Holm-Bonferroni correction in publications reporting on significance testing for all image modalities (A) and all analysis methods (B).

doi:10.1371/journal.pone.0110300.g008
Discussion

This systematic review investigated the use and performance of heterogeneity or texture quantification methods in radiological images for differentiation between tumor types, tumor grading, outcome prediction and treatment response monitoring. After a systematic literature search yielding 7351, 209 unique studies reported on heterogeneity as an imaging biomarker in tumor imaging. Since 2008, an increasing number of publications have reported on quantification of tumor heterogeneity. Since the present review is based on the existing literature, it reflects the modalities, heterogeneity analysis methods, and location of tumors that were investigated by the authors of the included studies. Because almost all of the included publications presented positive results, it should be noted that this literature probably contains an over presentation of modalities, heterogeneity analysis methods and tumor locations for which heterogeneity analysis seems to work.

Until 2006 most heterogeneity papers were based on US, whereas after 2007 there was an increase in the number of studies using MRI. During the present study period, NSM and SGLM were the most frequently used methods. Most of the papers focus on heterogeneity quantification to differentiate between tumor types, tumor grading or outcome prediction; however, the number of papers with the goal of response monitoring has recently increased. In tumor heterogeneity quantification, US is the most frequently used imaging modality for differentiation between tumor types, tumor grading and outcome prediction, and MRI is the most frequently used modality for treatment response monitoring. For monitoring of treatment response, NSM is the most frequently used method. To differentiate between tumor types and tumor grading, all methods are evenly distributed over all the modalities.

The performance of the heterogeneity features was mostly (67%) evaluated by classification experiments reporting performance measures such as accuracy, sensitivity, specificity and AUC. Papers reporting only on the results of the combination of texture features with other features were excluded from the analysis. Some authors selectively report on sensitivity without mentioning the specificity. The AUC is the preferred measure to report performance as it is more comprehensive compared to a measure based on a single threshold, such as accuracy. Only one paper reported an AUC of 0.5, all other papers reported higher values. This is most likely caused by publication bias: only the positive performance of heterogeneity features tend to reach the journals. Only 63% of the publications reporting classification results described the use of the cross-validation technique to limit the effect of over-fitting on the available data. We found no relation between the performance measures and the modality, or with the analysis method used. However, a negative correlation was found between the tumor-feature ratio and the AUC. When more tumors were available per selected feature, the AUC was lower. This correlation may be the result of overfitting of the data when fewer tumors per feature are available.

Publications using statistical hypothesis testing often did not perform a correction of the significance levels for multiple comparisons. For eight papers, due to missing information, a retrospective Holm-Bonferroni correction could not be performed by the authors. For 45% of the papers, the number of significant features decreased after the Holm-Bonferroni correction. We found no relation between the number of significant features after the Holm-Bonferroni correction and the modality or the analysis method used.

The number of prospective studies is small, i.e. only 13% of all studies. These latter studies are mainly based on MRI and report NSM features. Although the use of retrospectively collected data is necessary to develop, test and evaluate heterogeneity as a biomarker for differentiation between tumor types, tumor grading, outcome prediction and treatment response monitoring, the real test is to evaluate the performance of the developed features in a prospective study design. When using a retrospective study design, the criteria for the inclusion of cases are often not (or not clearly) described, so that the performance of the heterogeneity feature can be overestimated. Using a prospective study design, with clear inclusion criteria, the actual performance of heterogeneity features can be more reliably assessed.

Moreover, in most included studies, performance of the heterogeneity feature is evaluated without taking into account currently accepted clinical features, such as mean signal intensity, tumor size, tumor grade, or border regularity of a tumor. Some studies report only the combined classification performance of heterogeneity and clinical features. A large number of publications even use the mean signal intensity as a feature to estimate tumor heterogeneity, even though this is clearly not a heterogeneity measure (i.e., mean signal intensity does not measure intra-tumor heterogeneity). Based on these types of studies, it is not possible to evaluate the added value of heterogeneity to currently accepted clinical features. Whereas researchers are interested in the performance of the feature itself, clinicians are interested in the additional value of the feature compared with the currently available clinical biomarkers. Since the quantification of heterogeneity is usually more complex and computationally more costly than computing the mean intensity, the benefit of the added effort to characterize heterogeneity needs sufficient motivation. To enable the translation of imaging biomarkers from the research stage to clinical practice, future research should focus on studies investigating the additional value of the proposed heterogeneity biomarker compared with the established clinical markers. In this systematic review, comparison between the performance of different methods for a certain classification task was not possible due to the large variety in the datasets used and the classification tasks posed. The search for new and optimal (combinations of) heterogeneity features would benefit from developing reliable datasets (for different classification problems) that are available to the scientific community. Large well-defined datasets are a prerequisite for objective comparison of methods.

Future studies should have a design that takes the requirements from pattern recognition into account, i.e. a balanced number of subjects and features, cross-validation, independent test datasets, and a prospective study design. Satisfying these requirements will allow more reliable evaluation of the value of heterogeneity features.

Supporting Information

Figure S1 Numbers of publications for a specific imaging modality and analysis method. The supplementary EndNote files corresponding to the records for these publications (for each cell in the matrix separately) are publicly available. To download separate files just click on a cell of interest in the figure. (PDF)

Figure S2 The accuracy for different imaging modalities (A) and for different analysis methods (B) as a function of tumor-feature ratio in the classification experiments. The scatter plot shows each imaging modality and analysis method separately. Dotted line represents the ratio of 10 tumors per selected feature. Boxplot
representing distribution in AUC for imaging modality (C) and for analysis methods (D).

**Figure S3** The sensitivity for different imaging modalities (A) and for different analysis methods (B) as a function of tumor-feature ratio in the classification experiments. The scatter plot shows each imaging modality and analysis method separately. Dotted line represents the ratio of 10 tumors per selected feature. Boxplot representing distribution in AUC for imaging modality (C) and for analysis methods (D).

**Boxplot representing distribution in AUC for imaging modality (C) and for analysis methods (D).**

**(EPS)**

**References**

1. Fisher R, Pauzat L, Swanton C (2013) Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 107: 479–485.
2. Ng CK, Pemberton HT, Reis-Filho JS (2012) Breast cancer intratumor genetic heterogeneity: causes and implications. Expert Rev Anticancer Ther 12: 1021–1032.
3. Brown JR, DiGiovanna MP, Källlela B, Lannin DR, Rimm DL (2014) Quantitative assessment Ki-67 score for prediction of response to neoadjuvant chemotherapy in breast cancer. Lab Invest 94: 98–106.
4. Fasching PA, Heusinger K, Haererle I, Nikolai M, Heim A, et al. (2011) Ki67, chemotherapy response, and prognosis in breast cancer patients receiving neoadjuvant treatment. BMC Cancer 11: 496–498.
5. Szerlip NJ, Pedraza A, Chakravarty D, Azim A, McGuire J, et al. (2012) Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRα amplifications in glioblastoma defines subpopulations with distinct growth factor response. Proc Natl Acad Sci USA 109: 3041–3046.
6. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, et al. (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5: 1–8.
7. Hayes C, Padhani AR, Leach MO (2002) Assessing changes in tumour vascular function using dynamic contrast-enhanced magnetic resonance imaging. NMR Biomed 15: 154–163.
8. van Rijswijk CS, Geirnaerd MJ, Hogendoorn PC, Peters JL, van Goeverden F, et al. (2003) Dynamic contrast-enhanced MR imaging in monitoring response to isolated limb perfusion in high-grade soft tissue sarcoma: initial results. Eur Radiol 13: 1849–1858.
9. Pickard MD, Martinon DJ, Lowry M, Turnbull LW (2009) Prognostic value of pre-treatment DCE-MRI parameters in predicting disease free and overall survival for breast cancer patients undergoing neoadjuvant chemotherapy. Eur J Radiol 71: 491–505.
10. Briët DM, Silhey GS, Prinsin LR, Scher RI, Deshawit M (1997) Tumor hypoxia adversely affects the prognosis of carcinoma of the head and neck. Int J Radiat Oncol Biol Phys 38: 285–289.
11. Aerts HJ, Basink J, Oyen WJ, van Elmpft W, Folgering AM, et al. (2012) Identification of residual metabolic-active areas within NSCLC tumours using a pre-radiotherapy FDG-PET-CT scan: a prospective validation. Lung Cancer 75: 73–76.
12. Lambin P, Petit SF, Aerts HJ, van Elmpft W, Oeperie CJ, et al. (2010) The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. Radiother Oncol 96: 145–152.
13. PET Boost trial. Dose escalation by boosting radiation dose within the primary tumor on the basis of a pre-treatment FDG-PET-CT scan in stage IB, II and III NSCLC: a randomized Phase II trial. Available: www.clinicaltrials.gov.
14. Sinha S, Lucas-Quesada FA, DeBruhl ND, Sayar J, Farria D, et al. (1997) Multifrequency analysis of GD-enhanced MR images of breast lesions. J Magn Reson Imaging 7: 1016–1026.
15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 6: e1000097.
16. Alc C, Vermand J, Nissen ME (2012) Quantification of heterogeneity as a biomarker in tumour imaging: a systematic review. Available: http://wwwmetaxam.com/propro/full_docap/RecordID=3634 2015: 7328148.
17. Yang X, Knopp MV (2011) Quantifying tumor vascular heterogeneity with dynamic contrast-enhanced magnetic resonance imaging: a review. J Biomed Biotechnol 2011: 729084.
18. Asselin MC, O’Connor JP, Boellaard R, Thaker NA, Jackson A (2012) Quantifying heterogeneity in human tumours using MRI and PET. Eur J Cancer 48: 447–455.
19. Dvernul F, Yip CS, Ljungqvist G, Selmi M, Ng F, et al. (2012) Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging 3: 573–589.
20. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Statistics 6: 65–70.
21. Harbeck RO, Schumurmann K, Dinter J (1975) Textural features for image classification. IEEE Trans Syst Man Cybern 6: 610–621.
22. Amadason M, King R (1989) Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 19: 1264–1273.
23. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29: 51–59.
24. Mandelbrot BB (1983) The fractal geometry of nature. New York: W.H. Freeman. 468 p.
25. Smith TG Jr, Lange GD, Marks WB (1996) Fractal methods and results in cellular morphology—dimensions, lacunarity and multifractals. J Neurosci Methods 69: 123–136.
26. Fedg S, Naor J, Harley R, Aspin D (1984) Multiple resolution texture analysis and classification. IEEE Trans Pattern Anal Mach Intell 6: 518–525.
27. Tabachnick BG, Fidell LS (2013) Using multivariate statistics. Boston: Pearson Education. xxvi, 803 p.
28. Pekalska E, Duin RPW (2005) The dissimilarity representation for pattern recognition: foundations and applications. Hackensack, N.J.: World Scientific. xxvi, 607 p.
29. Young TY, Calvert TW (1974) Classification, estimation, and pattern recognition: American Elsevier Pub. Co. 366 p.
30. Acharya UR, Faust O, Seev SV, Molinari F, Garberoglio R, et al. (2011) Cost-effective and non-invasive automated benign and malignant thyroid lesion classification in 3D contrast-enhanced ultrasound using combination of wavelets and textures: a class of ThyroScan algorithms. Technol Cancer Res Treat 10: 371–380.
31. Acharya UR, Faust O, Seev SV, Molinari F, Suri JS (2012) ThyroScreen system: high resolution ultrasound thyroid image characterization into benign and malignant classes using novel combination of texture and discrete wavelet transform. Comput Meth Progr Biomed 107: 233–241.
32. Chang RF, Wu WJ, Moon WK, Chen DR (2003) Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture transform. Comput Meth Progr Biomed 107: 233–241.
33. Chang RF, Wu WJ, Moon WK, Chen DR (2003) Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture transform. Comput Meth Progr Biomed 107: 233–241.
34. Chang RF, Wu WJ, Moon WK, Chen DR (2003) Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture transform. Comput Meth Progr Biomed 107: 233–241.
35. Chang RF, Wu WJ, Moon WK, Chen DR (2003) Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture transform. Comput Meth Progr Biomed 107: 233–241.
40. Chen W, Giger ML, Bick U, Newstead GM (2006) Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. Med Phys 33: 2078-2087.

41. Ganeshes B, Ababei S, Young RC, Chavvin CR, Miles KA (2010) Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 10: 137–143.

42. Georgiadis P, Cavouros D, Kalazini I, Daskalakis A, Kagsidis GC, et al. (2008) Improving breast tumor characterization on MRI by probabilistic neural networks and non-linear transformation of textural features. Comput Meth Programs Biomed 89: 24–32.

43. Georgiadis P, Kostopoulou S, Cavouros D, Glotsoe D, Kalazini I, et al. (2011) Quantitative combination of volumetric MR imaging and MR spectroscopy data for the discrimination of meningiomas from metastatic brain tumors by means of pattern recognition. Magn Reson Imaging 29: 525–535.

44. Harrison L, Bastidar P, Eskola H, Jarvenpaa R, Pertovaara H, et al. (2009) Classification of MR images of non-Hodgkin lymphoma. Comput Med Biol 30: 519–524.

45. Kido S, Kuriyama K, Higashiya M, Kasagi T, Kuroda C (2002) Fractal analysis of small peripheral pulmonary nodules in thin-section CT: evaluation of the lung-nodule interfaces. J Comput Assist Tomogr 26: 573–578.

46. Klein HM, Klose KC, Eisele T, Brenner M, Ameling W, et al. (1993) [The diagnosis of focal liver lesions by the texture analysis of dynamic computed tomographs]. Klin Radiol 139: 10–15.

47. McNicoll-Grant MF, Hart EM, Wyrckoff N, Sayer JW, Goldin JG, et al. (1999) A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. Med Phys 26: 898–908.

48. Ng F, Koziarski R, Ganeshes B, Goh V (2013) Assessment of tumor heterogeneity by texture feature analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis? Eur J Radiol 82: 342–348.

49. O’Sullivan F, Roy S, Eary J (2003) A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. Biostatistics 4: 435–448.

50. Sun T, Wang J, Xu L, Liu F, et al. (2012) Comparative evaluation of support vector machines for computer-aided diagnosis of lung cancer in CT based on a multi-dimensional data set. Comput Med Programs Biomed 111: 519–524.

51. Quissens JM, Verbeek AM, Romijn RL, de Wolff-Rouendaal D, Oosterhuis JA (1991) Echographic differentiation of histological types of intraocular melanoma. Ultrasonic Med Biol 17: 127–130.

52. Way TW, Harijoji LM, Sahiner B, Chan HP, Cascade PN, et al. (2006) Computer-aided diagnosis of pulmonary nodules: CT scans: segmentation and classification using 3D active contours. Med Phys 33: 2323–2327.

53. Wu WJ, Moon WK (2008) Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features. Acad Radiol 15: 673–680.

54. Chen DR, Ho HC, Huang CH, Chang CY, Chang KY, et al. (2001) Ultrasound texture classification using run difference matrix. Ultrasound Med Biol 27: 1067–1076.

55. Downey K, Richards SF, Morgan VA, Giles SL, Artygalle AD, et al. (2013) Relationship between imaging biomarkers of stage I cervical cancer and poor-prognosis histologic features: quantitative histogram analysis of diffusion-weighted MR images. Am J Roentgenol 200: 514–520.

56. Dupont S, Roldan G, de Robles P, Adler D, McIntyre JB, et al. (2010) An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging. NeuroImage 49: 1398–1405.

57. Dumrongpisutikul N, Intrapiromkul J, Yousen DM (2012) Distinguishing between germinomas and pineal cell tumors on MR imaging. AJNR Am J Neuroradiol 33: 550–555.

58. Eary JF, O’Sullivan F, O’Sullivan J, Conrad EU (2008) Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. J Nucl Med 49: 1457–1459.

59. Eiat PA, Lechaux D, Gervais A, Rioux-Leclerc N, Franconi F, et al. (2001) Is texture analysis of sonographic images of the breast useful for the assessment of malignant potential? Eur Radiol 11: 169–171.

60. Eiat PA, Lechaux D, Gervais A, Rioux-Leclerc N, Franconi F, et al. (2001) Is texture analysis of sonographic images of the breast useful for the assessment of malignant potential? Eur Radiol 11: 169–171.

61. Eiat PA, Lechaux D, Gervais A, Rioux-Leclerc N, Franconi F, et al. (2001) Is texture analysis of sonographic images of the breast useful for the assessment of malignant potential? Eur Radiol 11: 169–171.

62. Embed KE, Nedergaard B, Nørræne T, Hansen D, Hald JK, et al. (2013) Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification of characteristic kinetic curves of breast lesions on DCE-MRI. Eur Radiol 23: 157–165.

63. Emhjuler L, Fredriksson J, Englund M, Agren G (1997) Weighted MR images. Am J Roentgenol 200: 314–320.

64. Facciorusso A, Salvatore L, Stabile F, Giorgio T, Pintauro F, et al. (2010) Automatic texture-based analysis in ultrasound imaging of ovarian masses. Ultrassall Med 34: 145–150.

65. Feit AE, Novak J, Rodriguez D, Auer DP, Clark CA, et al. (2013) MRI texture analysis in pediatric oncology: a preliminary study. Stud Health Technol Inform 190: 169–171.

66. Fruhwehl-Pallamar J, Czerny C, Holzer-Fruhwehl L, Nemec SF, Mueller-Mang C, et al. (2013) Texture-based and diffusion-weighted discrimination of parotid gland lesions on MR images at 3.0 Tesla. NMR Biomed 26: 1450–1459.

67. Ganesan B, Panayiotou E, Bumand K, Dzidarevic S, Miles K (2012) Tumor heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol 22: 796–802.

68. Ganesan B, Skogen K, Pernanen T, Coutroubis D, Miles K (2012) Tumor heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival. Clin Radiol 67: 156–174.
113. Jung SC, Cho JY, Kim SH (2012) Subtype differentiation of small renal cell carcinomas on three-phase MDCT: usefulness of the measurement of degree and heterogeneity of enhancement. Acta Radiol 53: 112–118.
114. Junto J, Sijbers J, De Backer S, Rajan J, Van Dyck D (2010) Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. J Magn Reson Imaging 31: 680–689.
115. Karahaliova A, Vassios K, Arkidis NS, Skapoulous S, Kanavos T, et al. (2014) Assessment of homogeneity and retest reliability in dynamic contrast-enhanced MRI for breast cancer diagnosis. Br J Radiol 83: 296–309.
116. Kidd EA, Grigby PW (2008) Intratumoral metabolic heterogeneity of cervical cancer. Clin Cancer Res 14: 5236–5241.
117. Kim SS, Kumazawa K, Higashiyama M, Kamoji T, Kuroda C (2003) Fractal analysis of internal and peripheral textures of small peripheral bronchogenic carcinomas in thin-section computed tomography: comparison of bronchio-vascular and nonbronchovascular cell carcinomas. J Comput Assist Tomogr 27: 66–71.
118. Kim DV, Kim JH, Nah SM, Park JW (2003) Pulmonary nodal detection using chest CT images. Acta Radiol 44: 252–257.
119. Kim KG, Cho SW, Min SJ, Kim JH, Min BG, et al. (2005) Computerized scheme for assessing ultrasonographic features of breast masses. Acad Radiol 12: 654–666.
120. Kim KG, Kim JH, Min BG (2001) Comparative analysis of texture characteristics of malignant and benign tumors in breast ultrasonograms. J Digit Imaging 14: 280–210.
121. Kögel L, Ring P, Thonning C, Henrichs O (1995) Texture analysis in quantitative MR imaging. Tissue characterization of normal brain and intracranial tumors at 1.5 T. Acta Radiol 36: 127–133.
122. Klein HM, Eisele T, Klose KG, Stans I, Brenner M, et al. (1996) Pattern recognition system for focal liver lesions using “crisp” and “fuzzy” classifiers. Invest Radiol 31: 6–10.
123. Kratzik C, Schuster E, Hainz A, Kuber W, Langmayr G (1988) Texture analysis—a new method of differentiating protrastatic carcinoma from protrastatic carcinoma. Acta Radiol 29: 497–500.
124. Kunzel C, Glaser F, Zuna I, Buhr HJ, Herfarth C (1994) Endorectal ultrasonography and computed b-scan texture analysis to assess sessile adenoma and small rectal carcinoma. Endoskopie Heute 7: 173–178.
125. Kuo WJ, Chang RF, Lee CC, Moon WK, Chen DR (2002) Retrieval technique for the diagnosis of solid breast tumors on sonogram. Ultrason Med Biol 28: 903–909.
126. Kuo WJ, Chang RF, Moon WK, Lee CC, Chen DR (2002) Computer-aided diagnosis of breast tumors with different US systems. Acad Radiol 9: 793–799.
127. Kurti K, Landboe S, Kalmo H, Valsson S (1995) MR classification of brain gliomas: value of magnetization transfer and conventional imaging. Magn Reson Imaging 13: 501–511.
128. Lai YC, Huang YS, Wang DW, Tso CM, Chou YH, et al. (2013) Computer-aided diagnosis for 3-d power Doppler breast ultrasound. Ultrason Med Biol 39: 555–567.
129. Larkin TJ, Canuto HG, Kettunen MI, Booth TC, Hu DE, et al. (2013) Analysis of image heterogeneity using 2D Minowski functions detects tumor margins to treatment. Med Phys 40: 198–207.
130. Lee CC, Shih CY (2010) Learning patterns of liver masses using improved RB networks. Biomedical Engineering - Applications, Basis and Communications 22: 137–147.
131. Leftihre F, Meunier M, Thibault F, Laugier P, Berger G (2000) Computer-assisted ultrasound b-scan characterization of breast nodules. Ultrasonic Med Biol 26: 1421–1428.
132. Li X, Yu L, Pirzalla A, Mc-Night T, Nelson SJ (2002) Analysis of the spatial characteristics of metabolic abnormalities in newly diagnosed glioma patients. J Magn Reson Imaging 16: 229–237.
133. Liao YY, Tsai PH, Li CH, Chang KJ, Kuo WH, et al. (2011) Classification of scattering media within benign and malignant breast tumors based on ultrasound texture-feature-based and Nakagami-parameter images. Med Phys 38: 2198–2207.
134. Liu F, Kornecki A, Shumilovich O, Gelnan N (2011) Optimization of time-to-peak analysis for differentiating malignant and benign breast lesions with dynamic contrast-enhanced MRI. Acta Radiol 10: 694–704.
135. Liu Y, Cheng HD, Hsu M, Chang YT, Tung YC, et al. (2012) Computer-aided diagnosis system for breast cancer diagnosis based on color Doppler flow imaging. J Med Syst 36: 3973–3982.
136. Lou YH, Mufat M, Das T, Liu X, Robson K, et al. (2012) Classification of MR tumor images based on Gabor wavelet analysis. J Med Biol Eng 32: 22–28.
137. Loren DE, Seghal CM, Ginsberg GG, Kochman ML (2002) Computer-assisted analysis of lymph nodes detected by EUS in patients with esophageal carcinoma. Gastrointest Endosc 56: 742–749.
138. Ma JH, Kim HS, Riu N, Cho KG (2010) Differentiation among glablastoma multiforme, solitary metastatic tumor, and lymphoma using whole-tumor histogram analysis of the normalized cerebral blood volume in enhancing and per enhancing lesions. Am J Neuroradiol 31: 1699–1706.
139. Maruyama H, Takahashi M, Sekimoto T, Kamesaki H, Shimada T, et al. (2012) Heterogeneity of microbubble accumulation: a novel approach to discriminate between well-differentiated hepatocellular carcinomas and regenerative nodules. Ultrasound Med Biol 38: 383–388.
Heterogeneity as Biomarker in Tumour Imaging: Systematic Review

211. Virmani J, Kumar V, Kalra N, Khandelwal N (2013) A comparative study of computer-aided classification systems for focal hepatic lesions from B-mode ultrasound. J Med Eng Technol 37: 292–306.

212. Virmani J, Kumar V, Kalra N, Khandelwal N (2013) SVM-based characterization of liver ultrasound images using wavelet packet texture descriptors. J Digit Imaging 26: 530–543.

213. Wang H, Gao XH, Jia ZW, Li HK, Liang ZG, et al. (2010) Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image. Eur J Radiol 80: 232–239.

214. Way TW, Sahiner B, Chan HP, Hadjiiski L, Cascade PN, et al. (2009) Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Med Phys 36: 3086–3098.

215. Willems JM, Turkheimer FE, Kenny LM, Abagyan EO (2013) Quantification of intra-tumour cell proliferation heterogeneity using descriptors of 18F fluorothymidine-positron emission tomography. Phys Med Biol 58: 187–309.

216. Wu WJ, Li SW, Moon WK (2012) Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. Comput Med Imaging Graph 36: 627–633.

217. Yan K, Yu Y, Tunney E, Baraldi R, Liao L (2012) Clinical study of a noninvasive multimodal sono-contrast induced spectroscopy system for breast cancer diagnosis. Med Phys 39: 1571–1578.

218. Yang X, Mrozek E, Lustberg M, Jia G, Sammet S, et al. (2012) Microcirculatory fraction (MCF(I)) as a potential imaging marker for tumor heterogeneity in breast cancer. Magn Reson Imaging 30: 1059–1067.

219. Yoshida H, Casalino DD, Keserci B, Coskun A, Ozurtk O, et al. (2003) Wavelet-packet-based texture analysis for differentiation between benign and malignant liver tumours in ultrasound images. Phys Med Biol 48: 3753–3755.

220. Yoshida H, Nappi J, MacEneaney P, Rubin DT, Dachman AH (2002) Computer-aided diagnosis scheme for detection of polyps at CT colonography. Radiographics 22: 961–975.

221. Yoshura K, Miwa K, Yuasa K, Tokumori K, Kanda S, et al. (1997) Ultrasonographic texture characterization of salivary and neck masses using two-dimensional gray-scale clustering. Dentomaxillofac Radiol 26: 332–336.

222. Zacharaki EI, Wang S, Chawla S, Soo Yoo D, Wolf R, et al. (2009) Classification and segmentation of intracardiac masses in cardiac tumor echocardiograms. Comput Med Imaging Graph 33: 332–336.

223. Zacharaki EI, Wang S, Chawla S, Soo Yoo D, Wolf R, et al. (2009) Classification and segmentation of intracardiac masses in cardiac tumor echocardiograms. Comput Med Imaging Graph 33: 332–336.

224. Zacharaki EI, Wang S, Chawla S, Soo Yoo D, Wolf R, et al. (2009) Classification and segmentation of intracardiac masses in cardiac tumor echocardiograms. Comput Med Imaging Graph 33: 332–336.

225. Zhang X, Kauriwa T, Fukushima K, Hara H, Tsubota T, et al. (2006) Application of artificial neural network for the classification of liver ultrasound images: validation in a large patient population. Ultrasound Med Biol 32: 1377–1382.

226. Zheng Y, Englander S, Balach S, Zacharaki EI, Fan Y, et al. (2009) STEP: spatiotemporal enhancement pattern for MR-based breast tumor diagnosis. Med Phys 36: 3192–3204.

227. Zhou S, Shi J, Zhu J, Gao Y, Wang R (2013) Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image. Biomed Signal Process Control 8: 688–696.

228. Zhu Y, Tan Y, Hua Y, Wang M, Zhang G, et al. (2010) Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. J Digit Imaging 23: 61–65.

229. Liu H, Motoda H (2008) Computational methods of feature selection; Liu H, Motoda H, editors. Boca Raton: Chapman & Hall/CRC.