Physica Medica 61 (2019) 44–51

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing

Alberto Traverso[a,c,*,1], Michal Kazmierski[a,1], Zhenwei Shi[a], Petros Kalendralis[a], Mattea Welch[c], Henrik Dahl Nissen[b], David Jaffray[c], Andre Dekker[a], Leonard Wee[a]

[a] *Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, The Netherlands*
[b] *Danish Colorectal Cancer Center South, Vejle Hospital, Vejle, Denmark*
[c] *Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Canada*

ARTICLE INFO

Keywords:
Magnetic resonance imaging
Diffusion weighted imaging
Apparent diffusion coefficient
Radiomic feature reproducibility
Locally advanced rectal carcinoma

ABSTRACT

Quantitative imaging features (radiomics) extracted from apparent diffusion coefficient (ADC) maps of rectal cancer patients can provide additional information to support treatment decision. Most available radiomic computational packages allow extraction of hundreds to thousands of features. However, two major factors can influence the reproducibility of radiomic features: interobserver variability, and imaging filtering applied prior to features extraction. In this exploratory study we seek to determine to what extent various commonly-used features are reproducible with regards to the mentioned factors using ADC maps from two different clinics (56 patients). Features derived from intensity distribution histograms are less sensitive to manual tumour delineation differences, noise in ADC images, pixel size resampling and intensity discretization. Shape features appear to be strongly affected by delineation quality. On the whole, textural features appear to be poorly or moderately reproducible with respect to the image pre-processing perturbations we reproduced.

1. Introduction

Neo-adjuvant chemoradiotherapy (NACRT) followed by total mesorectal excision (TME) is the accepted standard of care for locally advanced rectal cancer (LARC) due to conclusive evidence of superior clinical outcome [1–5]. However, TME is a highly invasive procedure leading to bowel and bladder complications, and its added value for “good responders” is currently being debated [6,7].

Magnetic Resonance Imaging (MRI) has flexibility for imaging anatomy, physiological parameters and biochemical function, through appropriate choice of pulse sequences. Diffusion-weighted imaging in MRI allows construction of 3D maps of apparent diffusion coefficient (ADC) of water molecules, that are promising markers of internal tumour pores and cellular interstices wherein water molecules can migrate [8]. A change in mean value of ADC has been shown to be associated with tumour response in a number of different cancers, including LARC [9,10]. Joye et al. showed that combined PET (Positron Emission Tomography) and MRI imaging parameters were strongly associated with pCR or near-pCR [11]. The above-mentioned results led to an active search for quantitative imaging biomarkers (radiomic features) that could have prognostic/predictive power to support indication for treatment.

Radiomics refers to computerized extraction of a large number of quantitative image metrics from medical images, that may reveal a deeper level of detail than is accessible to an unaided human eye, with the intent of defining tumour sub-types [12]. While radiomics has been successfully applied for clinical outcome predictions in Computed Tomography (CT) and Positron Emission Tomography (PET), its application to MRI is less advanced. Despite the chosen modality, recent publications showed the importance of evaluating radiomic features sensitivity with respect to several scenarios: different acquisition settings, inter-observer variability in tumour's delineations, choice of particular computational settings prior to features extraction (i.e. image pre-processing). The results affirm that different categories of radiomic features are, in different forms, affected by the abovementioned scenarios. For example, textural metrics have been shown to sensitively change their values when computed using different quantization. Trying to isolate a set of features which appear to be robust to all these factors is of interest. Again, most of the available work on this topic was carried on lung and head and neck cancers using CT or PET. [R1]

[*] Corresponding author at: Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands.

E-mail address: alberto.traverso@maastro.nl (A. Traverso).

[1] These authors equally contributed to the manuscript.

https://doi.org/10.1016/j.ejmp.2019.04.009
Received 3 December 2018; Received in revised form 11 April 2019; Accepted 12 April 2019
Available online 28 April 2019

However, Hu et al. [13] did demonstrate that volume-normalized features were more stable than not normalized features extracted from CT; while for MRI global textural descriptors showed more temporal stability than local-regional texture parameters [14].

In this exploratory study of ADC radiomic features, we seek to determine to what extent are various commonly-used features sensitive to inter-observer disagreements in tumour delineation and the application of digital image filter prior to radiomic feature extraction, which is an adopted procedure used in most radiomic studies.

2. Material and methods

2.1. Images

Ethical clearance was obtained for re-analysis of pre-radiotherapy LARC images collected between 2009 and 2012 by a Dutch radiotherapy clinic for inclusion in the THeragnostic Utilities for Neoplastic Diseases of the Rectum (THUNDER) clinical trial (NCT00969657, dataset described in Ref. [15]). A subset of 23 patients was retrospectively extracted from the THUNDER set having a pre-treatment diffusion-weighted imaging (DWI) examination at gradients of 0, 300 s/mm^2 and 1100 s/mm^2. ADC maps were constructed directly from the above field gradients in the Siemens (Erlangen, Germany) MR scanner console. A retrospective set of 33 LARC patients undergoing routine care were extracted with review board permission at a Danish radiotherapy clinic (population details for all the cohorts available in the Supplementary material). Images with the same DWI field gradients had been obtained using a Philips (Eindhoven, The Netherlands) MR scanner. ADC maps were then constructed using an in-house Matlab script (MathWorks, Natick, USA). [R13]: the ADC maps were calculated on a voxel-by-voxel basis using axial slices for all the cohorts.

The above datasets are hereafter referred to as the “THUNDER” and “CLINIC” cohorts, respectively. Key elements of the image acquisition settings are given in Table 1 for each cohort. Other than reconstructed slice thickness and pulse sequence repetition time, the imaging parameters were nominally closely matched across the two devices.

2.2. Gross tumour volume (GTV) delineation

GTVs were manually delineated via a standardized consensus method between the operators. Specifically, the ADC was overlaid with a constant false-colour lookup table over the 1100 s/mm^2 image. Some anatomical details were visible in the latter, and using these as a guide, an outline of the hyper-intense ADC region inside and adjoining the rectum was then drawn in by hand. On the THUNDER cohort, three observers, working independently, delineated the tumour on a Mirada (Mirada Medical, Oxford, UK) workstation. In the CLINIC cohort, two observers, working independently, delineated the tumour on an Oncentra External Beam (Elekta AB, Stockholm, Sweden) workstation. One common observer (author AT) delineated on both THUNDER and CLINIC cohorts. Observers (median experience, 4 years; range 1–10) were trained by a resident radiation oncologist to identify relevant normal and abnormal anatomical structures within the ADC maps. In addition, original CT scans with annotated lesions for all the patients were available to the observers, so that they could be guided in the delineations in the ADC maps. The median DICE for both the cohorts was 0.75 (range 0.6–0.90). At the end, delineations were exported into a single DICOM RT Structure Set file per patient. Each patient’s ADC map was also exported in standard DICOM format.

Table 1
MR-DWI image acquisition parameters for the THUNDER and CLINIC cohorts discussed in the text.

	Thunder	CLINIC
ManufACTURER	Siemens	Philips
scanner model	Avanto	Ingenia
FIELD MAGNITUDE	1.5 T	1.5 T
slice thickness	6 mm	4.6 mm
pixel spacing	1.98 mm	1.82 mm
echo time	79 ms	82.7 ms
DWI GRADIENTS	0, 300, 1100 s/mm^2	0, 300, 1100 s/mm^2
repetition time	4300 ms	2852.6 ms

2.3. Image pre-processing

Digital pre-processing on ADC maps was applied prior to extracting features. This was intended to test the sensitivity of histogram and textural features, since shape features in *PyRadiomics* are entirely independent of pre-processing. For each patient, a baseline radiomic feature value was calculated on the native (unprocessed) map. Subsequently, the native ADC map was altered using one digital image pre-processing operations at a time – (i) filtering, (ii) pixel dimension resampling and (iii) intensity value discretization. All pre-processing was performed using only the functions embedded within the open-source *PyRadiomics* library [16], which were themselves based on *SimpleITK* functions [17,18]. All the filters were applied in 3D. Mathematical details of the image pre-processing operations are provided in the Supplementary Materials.

2.4. Features extraction

Radiomic feature extraction was performed with *PyRadiomics*. The open-source *PyRex* extensions (*https://github.com/zhenweishi/Py-rex*) were used to manage the conversions of DICOM and DICOM RT Structure files to binary masks. A total of 70 radiomic features were extracted from each subject; 18 first-order (FO) features based on the intensity histogram, 13 shape metrics (SM), 23 features based on gray-level co-occurrence matrices (GLCM) and 16 features based on gray-level size-zone matrices (GLSZM). Mathematical definitions of these features are given on the *PyRadiomics* feature documentation page (https://pyradiomics.readthedocs.io/en/latest/features.html). Details used for computations, are specified in the Supplementary Materials. It is important to note that out of the 70 features, 6 features available in *PyRadiomics* are not defined in the IBSI (Image Biomarker Standardization Initiative), namely: Maximum 3D Diameter Column. Slice, and Row (SM); Total Energy (FO); GLCM Homogenity1/2. All the remaining features correspond to the definitions provided by IBSI.

2.5. Statistical analysis

Statistical analysis was performed in R Studio (v1.1.383), R (v 3.5.1) and Python (v3.6.4). A Concordance Correlation Coefficient (CCC) [19] was chosen as the reproducibility metric to evaluate the agreement of radiomic feature values in the perturbed image (with pre-processing filters, re-binning and resampling) with respect to the baseline feature values in the native ADC map. For each possible image pre-processing function, a CCC was computed for the feature value in the perturbed ADC relative to the native ADC map. The reported stability metric is the mean value of CCC over all observers in the combined THUNDER and CLINIC sets. For inter-observer dependence, we computed the Intraclass Correlation Coefficient (ICC) [20] across patients for each feature

We proposed that a feature was reproducible if CCC ≥ 0.85, in keeping with one of the most commonly used thresholds reported in the literature [21]. Moderately reproducible features were arbitrarily defined as 0.65 < CCC < 0.85. However, features with CCC ≤ 0.65 were deemed poorly reproducible. A threshold value of 0.85 was used also for the ICC values to define reproducibility.

To quantify the degree of reproducibility of features between the two datasets, the features were ordered by descending mean CCC and

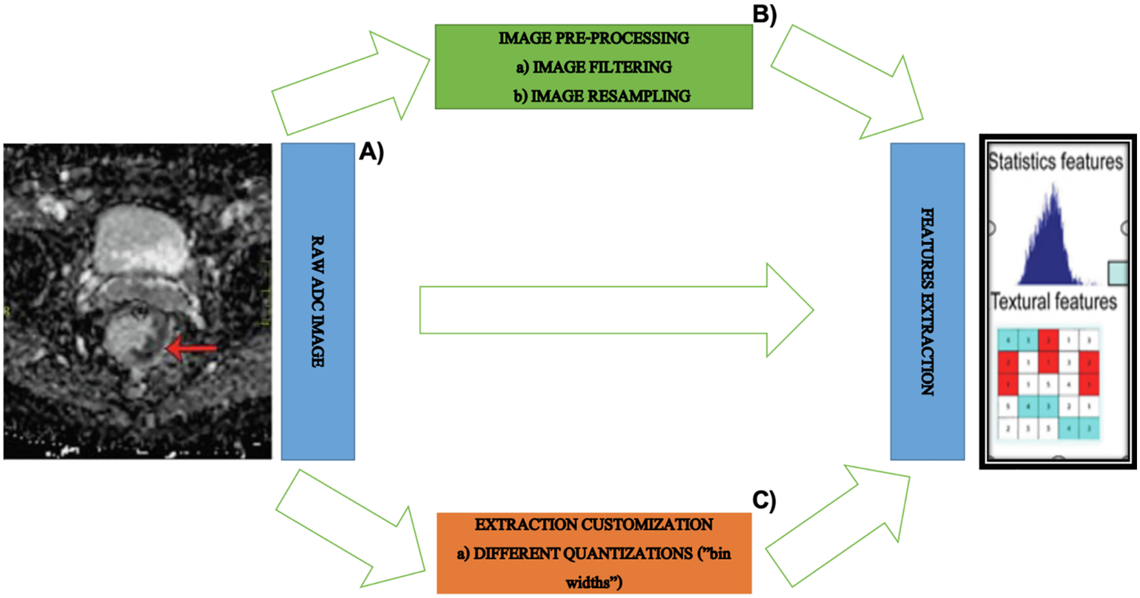

Fig. 1. Schematic representation of the workflow used in the analysis. Radiomic features are extracted from the GTVs annotated in the ADC maps. Different configurations were considered: (a) direct extraction from raw image, using default PyRadiomics settings; (b) customization of the extraction introducing different image pre-processing steps such as filtering or image resampling; (c) customization of the extraction, without modifying the original images, but considering different quantizations when computing features. The effects of (b) and (c) are then evaluating comparing differences in feature values with respect to (a) using concordance correlation coefficients.

compared using the Spearman Rank correlation coefficient [22]. Results were considered statistically significant if p-value < 0.05. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Spearman correlation at least as extreme as the one computed from these datasets. Fig. 1 proposes a sketch representation of the workflow used for the experiments.

3. Results

3.1. Inter-observer dependence

The overall sensitivity of feature types with respect to differences in manual GTV delineations were compared using the ICC metric. A box and whisker boxplot summarising the median ICC and its distribution for four feature types is given ins Fig. 2. Among the FO and GLCM feature types from the native ADC maps, the median ICC was consistently high in both THUNDER and CLINIC datasets. Major divergences appear for the GLSZM and SM feature types, with respect to the persons performing the delineations in the THUNDER and CLINIC datasets, respectively, such that GLSZM and SM features appeared more reproducible in the latter. There was significant spread in ICC for every feature type, so even within a related group of features certain individual features are much less sensitive to delineation differences than others.

3.2. Effect of resampling with interpolation

A heatmap of CCC ranges with respect to axial pixel dimension resampling is given as Fig. 3. At a glance, it is clear to see that the FO features are generally reproducible with respect to scale changes in pixel dimensions. The single FO feature that falls below CCC of 0.65 relative to the native ADC, after perturbation, happens to be "Energy" in this study. The GLCM features are moderately reproducible with resampling, since many features retain good or moderate reproducibility over a wide range of resampling. The majority of GLSZM features are, on the whole, poorly reproducible.

3.3. Effect of intensity value discretization

A heatmap of CCC ranges with respect to changes in the width of discrete intensity "bins" is given as Fig. 4. The overall trend here, once again, is that FO features (except for Kurtosis and Skewness) are generally reproducible over a wide range of intensity discretization bin widths; however, nearly all of the GLCM and GLSZM features are poorly reproducible. One texture feature – "GLSZM Gray Level Non-Uniformity" – could be a potential feature with good to moderate reproducibility with respect to image intensity discretization. However, previous studies pointed out the strong correlation between this feature and tumour volume. In the Supplementary material a list of most reproducible features is supplied.

3.4. Effect of applying digital image filters

A heatmap of CCC ranges with respect to application of different types of digital image filters is given as Fig. 5. In regard to feature types, FO features seem to be largely reproducible after additive Gaussian noise. The overall picture is more mixed with curvature flow, Laplacian and Gaussian smoothing filters, but it generally holds that GLCM and GLSZM feature types are poorly reproducible.

4. Discussion

We have used an ICC metric to examine the overall reproducibility of radiomic features with respect to tumour (GTV) delineation differences between groups of observers. Overall, we find that FO and GLCM feature types are less sensitive to manual delineation differences, but within each feature type a wide spread of ICCs are observed. We hypothesize that the experience level of an observer plays a key role in feature reproducibility, since we observe the median and distribution of feature ICC values appear more consistent across all feature types in the CLINIC set than compared to the THUNDER set. However, it is well known that interobserver variability increases as the quality of the image decreases. In fact, agreement between clinicians delineating on Computed Tomography (CT) is usually larger than on MRI or ADC, due to higher signal to noise ratio [23,24]. To improve agreement, studies

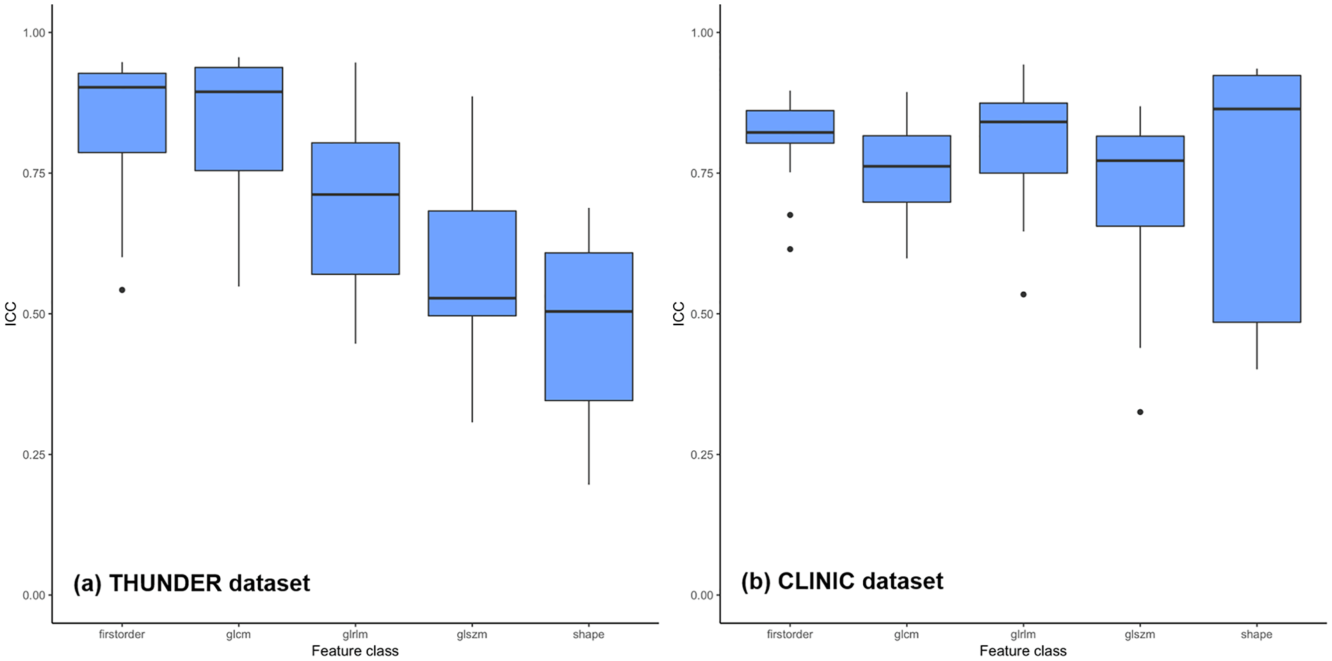

Fig. 2. Box and whisker plots of intraclass correlation coefficient (ICC) grouped by type of feature – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM) and shape metrics – for the (a) THUNDER and (b) Danish CLINIC datasets. The solid bars represent the median. The upper and lower edges of the boxes represent the upper and lower quartiles of the ICC distribution, respectively.

suggest defining strong protocols for delineations, and training of the observers to strictly follow the mentioned protocols. Qualitatively, taking both datasets into account, the overall trend for increasing risk of feature group irreproducibility appears to be – FO (least risk), GLCM, GLSZM and SM (greatest risk). This hypothesis finds some support in recent literature, where a recent study by van Heeswijk et al. [25] was able to identify a FO feature that was reproducible when a fast approximate delineation was used in place of a time-consuming precise delineation of a rectal tumour on ADC maps. However, that specific study did not consider other feature groups, except for a subset of FO features.

We also examined the reproducibility of types of radiomic features when a range of different image pre-processing operations were applied prior to feature extraction. We used a CCC metric to compare the feature value in the processed ADC map versus the same feature value in the native (unperturbed) ADC map. We have detected the overall trend that FO feature types were robust with respect to many of these perturbations, but GLCM and GLSZM were in general sensitive to such pre-processing. It is not surprising that FO features were less impacted by image pre-processing than textural features. In fact, FO features can be considered as global statistical descriptors, while textural features provide a local measurement by looking at particular patterns inside gray values. Being a local measurement, any image pre-processing that alters the local values, or the matrixes used for computation of TA, can produce values that are much different than the features computed on the original image. For example, a study [26] performed on a dedicated texture phantom for radiomics studies, showed TA features to be very sensitive to the bin width chosen for computation.

In totality, our results suggest that overall global intensity-based descriptors (such as the FO type) may be more tolerant to differences in GTV delineation accuracy, pixel dimensions, noise level and image-enhancing digital filters compared to textural features such as GLCM and GLSZM. These results were found to be consistent across the two different cohorts ($p < 0.01$). The found results are in line with a recent study [21] proposing a qualitative synthesis of 41 studies investigating the repeatability and reproducibility of radiomic features. From the analysis, textural features were found to be more sensitive than FO features with respect to inter-observer variability and image processing. However, the analysis also revealed the lack of a consensus. Furthermore, it shows that results could depend on the modality or the anatomical site considered. Unfortunately, due to the lack of literature investigating this topic for rectal cancers in MRI, it is not possible to have a quantitate meta-analysis. Nevertheless, as this study also shows, it becomes fundamental to report the exact details used for the computations prior to features extraction.

It is important to note that we do not make a claim about the potential predictive power of feature types, nor is it in the scope of this study to identify any set of features as more preferable than others. The CCC metrics show that image pre-processing has the potential to strongly change the value of some radiomic features relative to the same feature value in the unperturbed native image, but the data cannot substantiate whether this change is leading to better or worse predictive performance in the final model. Also, as pointed out in Ref. [27], when considering the prognostic/predictive power of radiomic features, their correlation with accepted clinical factors (such as for example tumour extension) should be considered. This is to avoid redundant information that might increase the risk of overfitting, while only features that provide additional information besides other predictors should be kept [28].

The purpose of this study is to emphasize that differences in the steps leading up to feature extraction could negatively affect the wider generalizability of any given model developed using radiomic features as a signature.

For instance, if it is known in advance that the intended application of a radiomic signature might include a wide range of pixel dimensions, it may be preferable to prioritize features whose values do not change greatly as a function of pixel size. Alternatively, if a particular radiomics signature uses a specific image-enhancing digital filter, it is almost certain that the exact same digital filter will be needed to obtain reasonable validation results, particularly if complex textural features are part of the radiomic signature being validated. Finally, it is important to verify the correlation between features.

Our cross-institutional dataset used in this investigation did not allow us to investigate additional aspects of ADC reproducibility. For

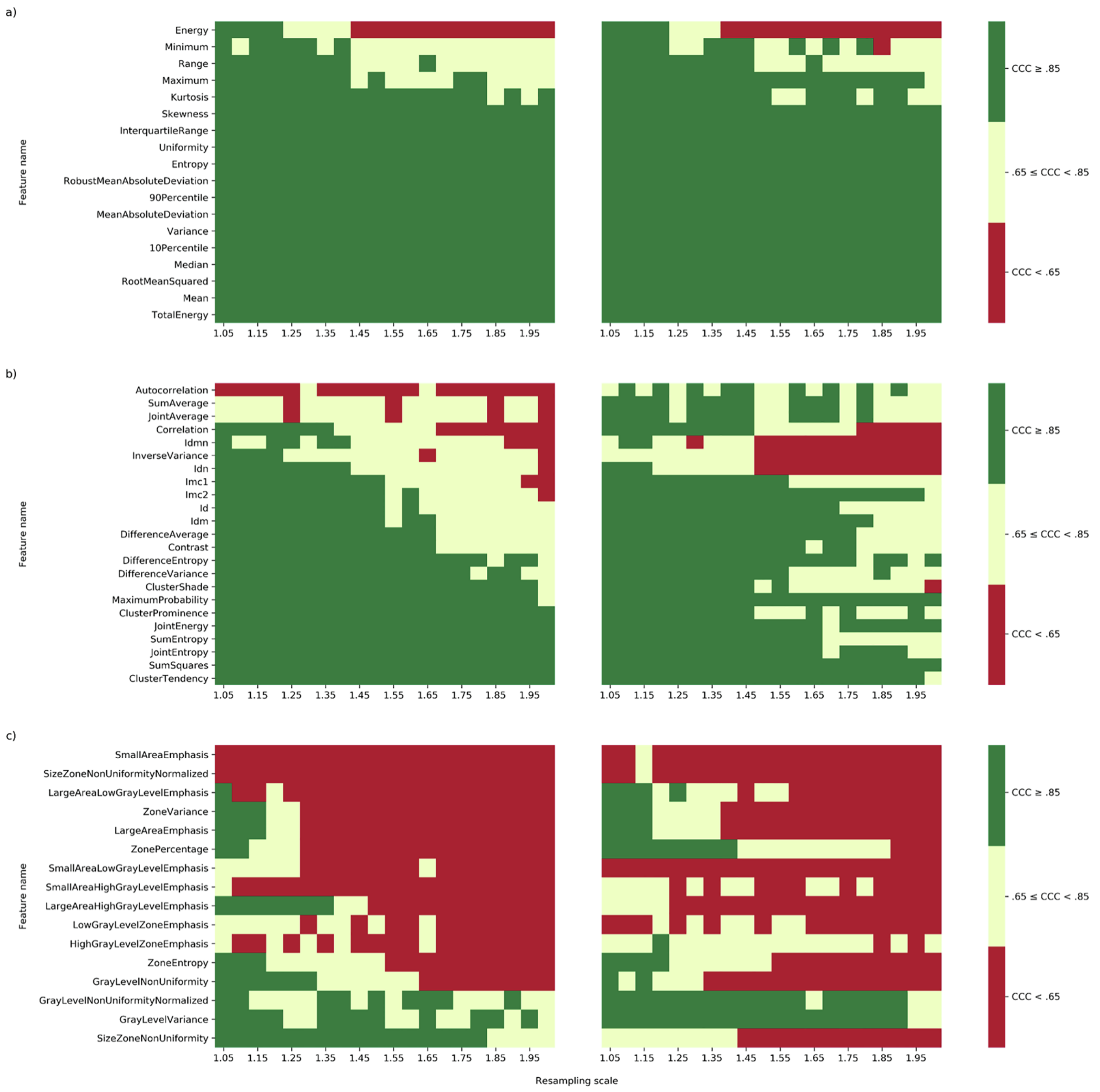

Fig. 3. Heatmap defining ranges of concordance correlation coefficient (CCC). The perturbation introduced is resampling of the pixel dimensions in the axial plane with interpolation, the magnitude of which is shown along the horizontal axis. Each row in the image corresponds to a particular feature within one of the feature types – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM). Shape metrics are not evaluated, because they are independent of resampling of the image pixel in the pyradiomics implementation. Results are shown on the left side for the CLINICAL cohort and on the right side for the THUNDER dataset.

example, it is well known that intensity in MR images may drift significantly over time. We did not have the data to examine temporal stability of ADC feature values, though this has been investigated by Newitt et al. [29] for breast tumours. Here, we used only THUNDER imaging series that were a nominal match of the field gradients available from the Danish CLINIC dataset. As pointed out by others, radiomic features may also depend on the number of unique DWI gradients and the magnitude of those gradients used when generating an ADC map [30,31]. As an additional limitation, we considered in our experiment the preliminary example of radiomic feature sensitivity with respect to the introduction of gaussian noise and the application of gaussian blurring to possibly reduce the noise. This was meant a) to test features' behavior in an 'extreme situation', considering that ADC maps already present a relevant intrinsic level of noise; b) verifying the sensitivity of features with respect to one possible de-noising technique. Further studies are needed to investigate the impact of noise in features' reproducibility.

Bologna et al. [32] further suggests that ADC feature reproducibility will depend on the region of the body being examined. This suggests that repeatability and reproducibility should be considered early in the radiomic model development process, by way of *a priori* feature selection. This would lead to better generalizability in external validation.

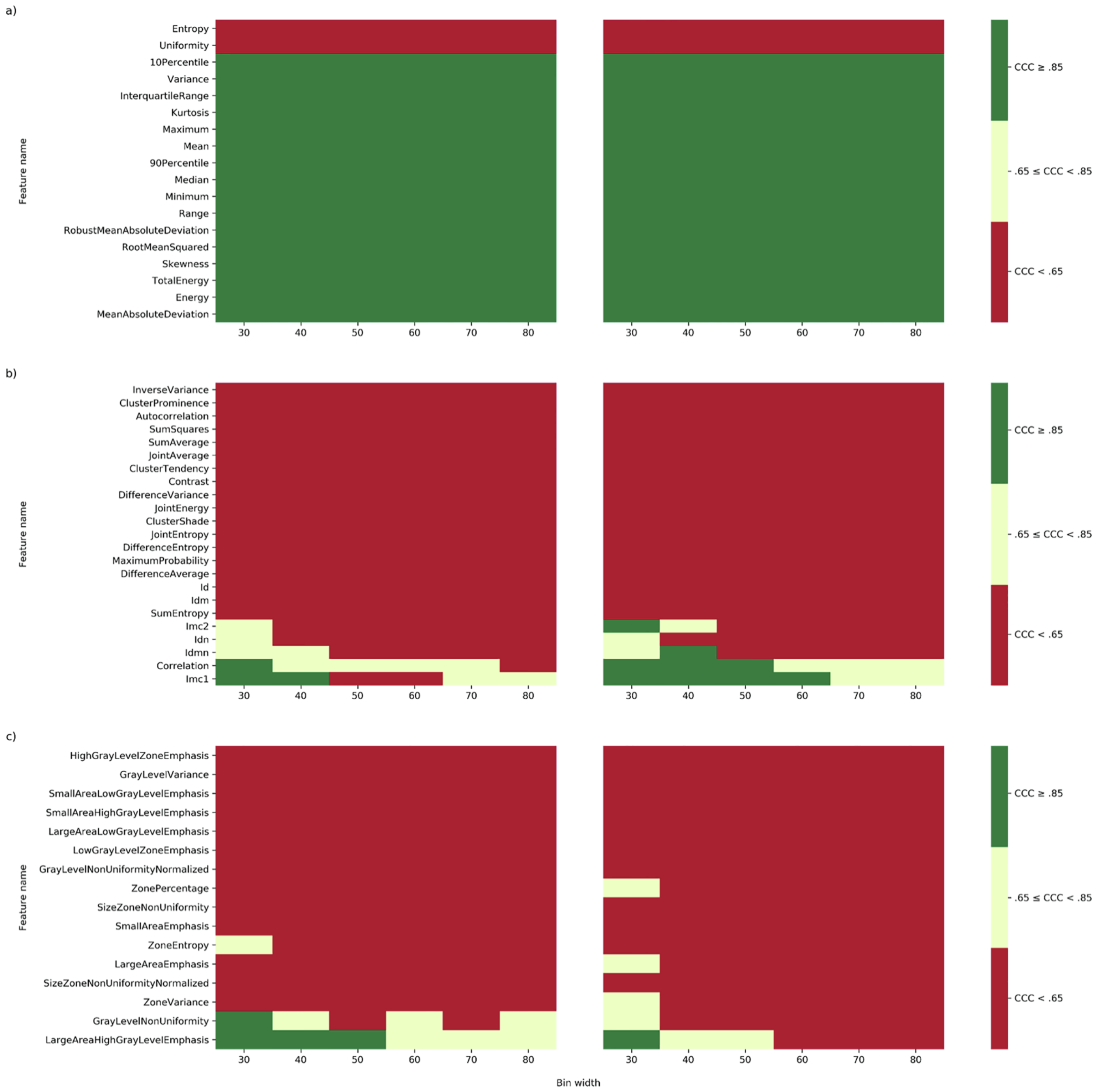

Fig. 4. Heatmap defining ranges of concordance correlation coefficient (CCC). The perturbation introduced is the discretization bin width for the image intensity values, the magnitude of which is shown along the horizontal axis. Each row in the image corresponds to a particular feature within one of the feature types – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM). Shape metrics are not evaluated, because they are independent of intensity discretization in the pyradiomics implementation. Results are shown on the left side for the CLINICAL cohort and on the right side for the THUNDER dataset.

Our future plans include the extension of our study to additional MRI sequences, such as T1 or T2 weighted imaging, which are often used as standard imaging for pelvic malignancies. In particular, we would like to verify if our results can be validated on different modalities, but within the same anatomical site. This will provide us with an initial evaluation of the sensitivity of radiomics features and related imaging pre-processing as a function of different modalities.

5. Conclusions

Evidence in literature clearly points towards the need to evaluate reproducibility of radiomic features derived on ADC maps. In this work, we demonstrated that – generally speaking – the mathematically simpler features, such as those derived from intensity distribution histograms, are less sensitive to manual tumour delineation differences, noise in ADC images, pixel size resampling and intensity discretization. Shape features appear to be strongly affected by delineation quality, and the expertise among groups of observer plays a role. On the whole, GLCM and GLSZM features appear to be poorly or moderately reproducible with respect to the image pre-processing perturbations we reproduced. Further studies are required to elucidate the role of diffusion gradients and temporal stability of DWI scans in order to develop the role of radiomic analysis in supporting treatment response monitoring in locally advanced rectal cancer.

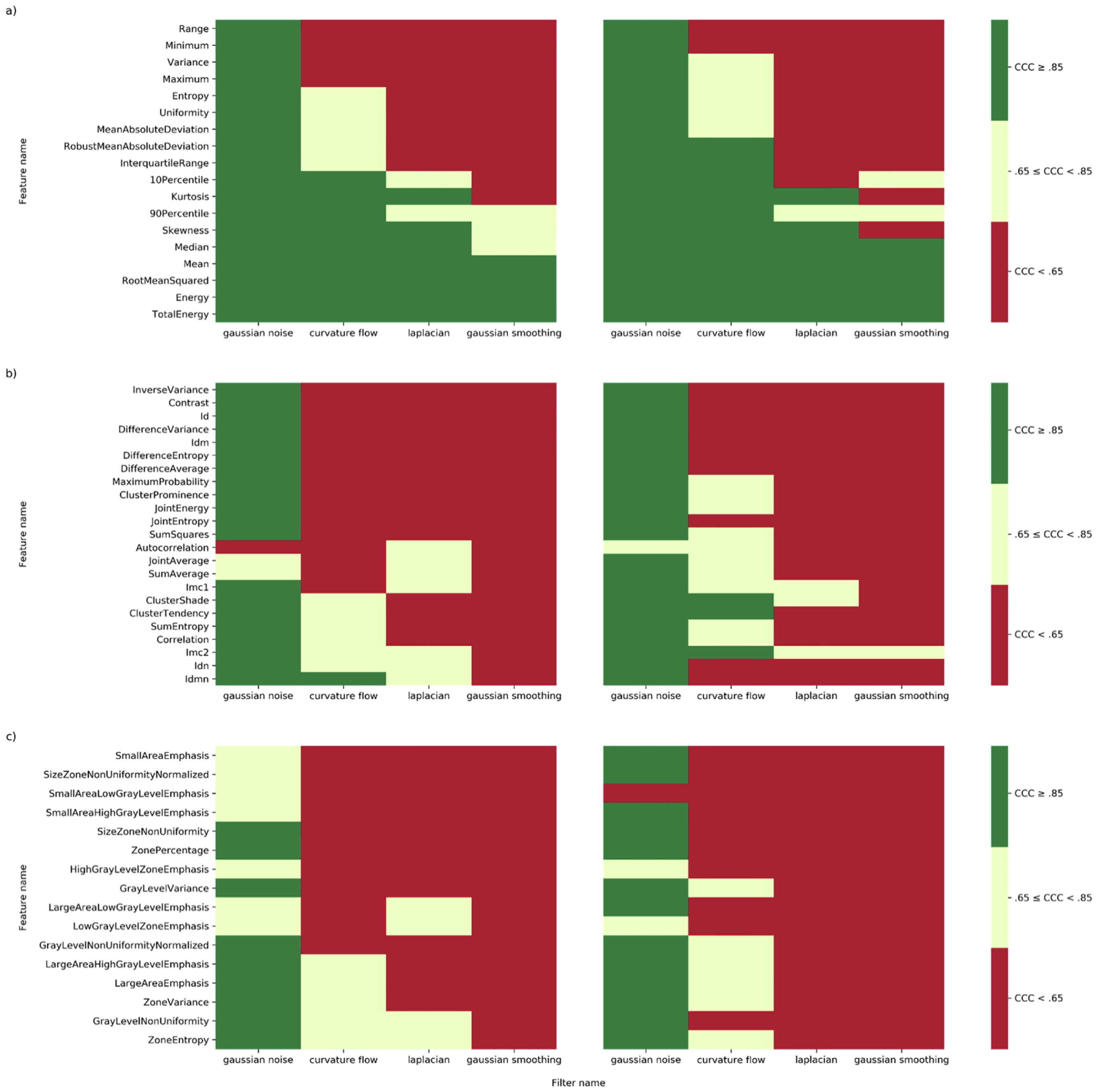

Fig. 5. Heatmap defining ranges of concordance correlation coefficient (CCC). The perturbations introduced are four different digital image manipulation filters as described in the main text, which is denoted along the horizontal axis. Each row in the image corresponds to a particular feature within one of the feature types – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM). Shape metrics are not evaluated, because they are independent of digital filtering in the pyradiomics implementation. Results are shown on the left side for the CLINICAL cohort and on the right side for the THUNDER dataset.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2019.04.009.

References

[1] MacFarlane JK, Ryall RD, Heald RJ. Mesorectal excision for rectal cancer. Lancet Lond Engl 1993;341:457–60.
[2] Sauer R, et al. Preoperative versus postoperative chemoradiotherapy for rectal cancer. N Engl J Med 2004;351:1731–40.
[3] Gérard J-P, et al. Preoperative radiotherapy with or without concurrent fluorouracil and leucovorin in T3–4 rectal cancers: results of FFCD 9203. J Clin Oncol Off J Am Soc Clin Oncol 2006;24:4620–5.
[4] Bosset J-F, et al. Chemotherapy with preoperative radiotherapy in rectal cancer. N Engl J Med 2006;355:1114–23.
[5] Maas M, et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. Lancet Oncol 2010;11:835–44.
[6] Plummer JM, Leake P-A, Albert MR. Recent advances in the management of rectal cancer: no surgery, minimal surgery or minimally invasive surgery. World J Gastrointest Surg 2017;9:139–48.
[7] Ma B, Xu Q, Song Y, Gao P, Wang Z. Current issues of preoperative radio(chemo) therapy and its future evolution in locally advanced rectal cancer. Future Oncol Lond Engl 2017;13:2489–501.
[8] Bonekamp S, Corona-Villalobos CP, Kamel IR. Oncologic applications of diffusion-weighted MRI in the body. J Magn Reson Imaging JMRI 2012;35:257–79.
[9] Heijmen L, et al. Tumour response prediction by diffusion-weighted MR imaging: ready for clinical use? Crit Rev Oncol Hematol 2012;83:194–207.
[10] Amodeo S, et al. MRI-based apparent diffusion coefficient for predicting pathologic

response of rectal cancer after neoadjuvant therapy: systematic review and meta-analysis. AJR Am J Roentgenol 2018;211:W205–16.

[11] Joye I, et al. Quantitative imaging outperforms molecular markers when predicting response to chemoradiotherapy for rectal cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol 2017;124:104–9.

[12] Aerts HJWL, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.

[13] Hu P, et al. Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget 2016;7:71440–6.

[14] Gourtsoyianni S, et al. Primary rectal cancer: repeatability of global and local-regional MR imaging texture features. Radiology 2017;284:552–61.

[15] van Stiphout RGPM, et al. Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: a multicentric prospective study with external validation. Radiother Oncol 2014;113:215–22.

[16] van Griethuysen JJM, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7.

[17] Lowekamp BC, Chen DT, Ibanez L, Blezek D. The design of SimpleITK. Front Neuroinformatics 2013;7.

[18] Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. J Digit Imaging 2018;31:290–303.

[19] Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255.

[20] Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3–11.

[21] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 2018;102:1143–58.

[22] Zar JH. Significance testing of the spearman rank correlation coefficient. J Am Stat Assoc 1972;67:578–80.

[23] Weltens C, et al. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. Radiother Oncol J Eur Soc Ther Radiol Oncol 2001;60:49–59.

[24] Buijsen J, et al. FDG-PET-CT reduces the interobserver variability in rectal tumor delineation. Radiother Oncol J Eur Soc Ther Radiol Oncol 2012;102:371–6.

[25] van Heeswijk MM, et al. Measuring the apparent diffusion coefficient in primary rectal tumors: is there a benefit in performing histogram analyses? Abdom Radiol NY 2017;42:1627–36.

[26] Larue RTHM, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol 2017;56:1544–53.

[27] Welch ML, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol J Eur Soc Ther Radiol Oncol 2018. https://doi.org/10.1016/j.radonc.2018.10.027.

[28] Hatt M, 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.

[29] Newitt DC, et al. Test-retest repeatability and reproducibility of ADC measures by breast DWI: results from the ACRIN 6698 trial. J Magn Reson Imaging JMRI 2018. https://doi.org/10.1002/jmri.26539.

[30] Newitt DC, et al. Multisite concordance of apparent diffusion coefficient measurements across the NCI quantitative imaging network. J Med Imaging Bellingham Wash 2018;5. 011003.

[31] Chen L, et al. Diffusion-weighted imaging of rectal cancer on repeatability and cancer characterization: an effect of b-value distribution study. Cancer Imaging Off Publ Int Cancer Imaging Soc 2018;18:43.

[32] Bologna M, et al. Assessment of stability and discrimination capacity of radiomic features on apparent diffusion coefficient images. J Digit Imaging 2018;31:879–94.