Investigating the impact of the CT Hounsfield unit range on radiomic feature stability using dual energy CT data

Avishek Chatterjee^{a,b,*}, Martin Vallières^{a,c}, Reza Forghani^d, Jan Seuntjens^a

^a McGill University, Medical Physics Unit, Montreal, Canada

^b Maastricht University, Dept. of Precision Medicine, Maastricht, The Netherlands

^c Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada

^d McGill University Health Centre, Dept. of Radiology, Montreal, Canada

ARTICLE INFO

Keywords:

Radiomics

CT

Feature stability

Replicability

ABSTRACT

Purpose: Radiomic texture calculation requires discretizing image intensities within the region-of-interest. FBN (fixed-bin-number), FBS (fixed-bin-size) and FBN and FBS with intensity equalization (FBNequal, FBSequal) are four discretization approaches. A crucial choice is the voxel intensity (Hounsfield units, or HU) binning range. We assessed the effect of this choice on radiomic features.

Methods: The dataset comprised 95 patients with head-and-neck squamous-cell-carcinoma. Dual energy CT data was reconstructed at 21 electron energies (40, 45,... 140 keV). Each of 94 texture features were calculated with 64 extraction parameters. All features were calculated five times: original choice, left shift (-10/-20 HU), right shift (+10/+20 HU). For each feature, Spearman correlation between nominal and four variants were calculated to determine feature stability. This was done for six texture feature types (GLCM, GLRLM, GLSZM, GLDZM, NGTDM, and NGLDM) separately. This analysis was repeated for the four binning algorithms. Effect of feature instability on predictive ability was studied for lymphadenopathy as endpoint.

Results: FBN and FBNequal algorithms showed good stability (correlation values consistently >0.9). For FBS and FBSequal algorithms, while median values exceeded 0.9, the 95% lower bound decreased as a function of energy, with poor performance over the entire spectrum. FBNequal was the most stable algorithm, and FBS the least.

Conclusions: We believe this is the first multi-energy systematic study of the impact of CT HU range used during intensity discretization for radiomic feature extraction. Future analyses should account for this source of uncertainty when evaluating the robustness of their radiomic signature.

1. Introduction

Radiomics converts medical images to mineable data which may be used to predict various clinical endpoints or outcomes [1]. Currently, commonly used imaging modalities are CT, MR, and PET. Radiomics has been part of the research lexicon since 2012 [2,3], and the field has burgeoned ever since. However, clinical translation has been slow to non-existent. A crucial barrier for the translation of radiomics to the clinical setting is the concern regarding replicability of published results. One approach towards overcoming these hurdles is to standardize the image acquisition as well as various parts of the typical radiomic workflow [1,4].

The workflow for handcrafted radiomics starts with image acquisition, followed by image segmentation to identify the region of interest

(ROI). Feature extraction from the identified ROI (typically the entire tumor, but possibly sub-regions of the tumor, also known as *habitats*, including tumor periphery, or even adjacent tissues depending on the pathology of interest) is next, and the final step is model-building and testing. Each process consists of sub-processes, and all of these need to be examined, optimized, and standardized to ensure replicability. In previous work [5], we have laid out recommendations regarding how to improve model-building and testing, and the adaptation of our recommendations would help standardize this step. In this work, we tackle an overlooked aspect of the feature extraction step and show that it is remarkably consequential in reality.

Feature extraction is perhaps the easiest part of the workflow to standardize, since it is entirely software-based and requires no user interaction. The pre-eminent effort in this direction is arguably the

* Corresponding author.

E-mail address: a.chatterjee@maastrichtuniversity.nl (A. Chatterjee).

<https://doi.org/10.1016/j.ejmp.2021.07.023>

Received 4 January 2021; Received in revised form 14 July 2021; Accepted 16 July 2021

Available online 27 July 2021

1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

Image biomarker standardisation initiative (IBSI) [6]. Image biomarkers can broadly be classified as texture features (derived from texture matrices that quantify the relationship between neighboring image voxels) and non-texture features (an umbrella term for intensity-based statistical features, intensity histogram-based features, intensity-volume histogram-based features, morphological features, and local intensity features), although it should be noted that the exact definition of texture features varies in the literature, particularly in the medical literature [7].

With few exceptions (e.g., morphological features and intensity-based statistical features), discretization of image intensities inside a ROI is required to calculate handcrafted features, in particular, texture features. Discretization may also aid noise-suppression. Two common approaches to discretization are a fixed number of bins (FBN), or bins of a fixed bin size (FBS). The choice of binning method (or the number of grey levels or the voxel spacing for interpolation) can be thought of as a hyper-parameter, and the user may either make a specific decision, or calculate features using all possible combinations of hyper-parameters.

During discretization, a crucial choice is the range of voxel intensity values over which the binning will be performed, a quantity named *re-segmentation range* by the IBSI. While this is not relevant for raw MR imaging data, where the intensity values are arbitrary, this is an important consideration for CT, where the reconstructed intensity values represent Hounsfield Units (HU) and have a physical meaning linked to the material or electron density of the reconstructed voxels. In this work, using Dual Energy CT (DECT) data from head-and-neck cancer patients, we study the impact of this choice of re-segmentation range on texture features. There are two central aims of this study: (1) to evaluate if feature robustness is dependent on the choice of discretization algorithm, and (2) to investigate if this robustness is dependent on CT electron energy, and if the findings change with energy.

2. Materials & methods

2.1. Description of clinical dataset

The dataset consisted of 95 patients with biopsy-proven mucosal head-and-neck squamous cell carcinomas (HNSCCs) below the hard palate including tumors arising from the oral cavity, oropharynx, hypopharynx, and larynx. All patient scans were obtained with a pre-defined imaging protocol using the same 64-slice dual-energy scanner (GE Discovery CT750HD) using the same technique after administration of 80 mL of iopamidol (Isovue 300®, BRACCO Diagnostics Inc.), injected at a rate of 2 mL/s, and the patient was scanned after a delay of 65 s. The dual energy CT data for each patient was reconstructed at 21 virtual monochromatic image (VMI) energy levels in 5 keV increments (40–140 keV). The ROI (tumor) was delineated by a radiology fellow, and then verified (and amended, if necessary) by a sub-specialty trained head-and-neck radiologist. For each case, an ROI was manually drawn on the image set where the tumor was best seen, typically the 40 keV VMIs, in conjunction with VMIs at other energies if needed [8]. The contouring was done using OsiriX MD (Bernex, Switzerland). Any tumor slices with imaging artifacts were removed from the study. If this resulted in a tumor getting separated into two or more sub-volumes, the largest sub-volume was retained for analysis. This was done to avoid interpolating over missing slices, as such interpolations are likely to be unreliable for tumors that are morphologically heterogeneous, have complex shapes and change noticeably from one CT slice to the next, as commonly encountered with HNSCC. For the sub-studies related to outcome prediction, the outcome of interest was the presence or absence (binary) of locoregional lymphadenopathy; of the 95 patients, 87 were eligible, of which 44 were node-positive.

2.2. Description of radiomic features

The radiomic features used were those defined by the IBSI. There were 76 non-texture features, and 94 texture features. Each texture feature was calculated in 64 ways by using all combinations of the following hyper-parameters: 4 voxel sizes (1 mm, 2 mm, 3 mm, 4 mm), 4 binning algorithms (FBN, FBNequal, FBS, FBSequal, their meaning explained in Section 2.4 gray level discretizations (8, 16, 32, 64 or 6.25 HU, 12.5 HU, 25 HU, 50 HU for FBN and FBS, respectively). Thus, there were 6092 features ($76 + 94 \times 64$) for each energy. The features were calculated on the 3D tumor volume (or largest sub-volume, for images with artifacts) using in-house code compatible with IBSI and written in MATLAB R2017b (MathWorks, Natick, MA). The code is available on GitHub (<https://github.com/mvallieres/radiomics-develop>).

2.3. Method to choose the HU window or re-segmentation range

For each patient, the central tumor slice was selected. For each of the 21 energies, the HU values within the ROI were extracted. This resulted in 1995 (= 95×21) vectors of HU values. These vectors were concatenated into a single vector, and the result was binned into a histogram. The window range was required to encompass a sufficient dynamic range ($\sim 10^3$). A window size of 400 HU is desirable, as it permits easier comparison between FBN and FBS algorithms.

2.4. Further details about feature extraction workflow

Two image processing steps are relevant to this study: re-segmentation and discretization. Re-segmentation may be performed to remove voxels from the intensity mask that fall outside of a specified range. An example is the exclusion of voxels with Hounsfield Units indicating air and bone tissue in the tumor ROI within CT images. When a re-segmentation range is defined by the user (e.g., a specified HU window as in the current study), it is used for the calculation of radiomic features for all ROIs, effectively excluding some imaging intensities from the analysis. Next, discretization of image intensities inside the ROI is required. For FBN, the number of intensity bins gets fixed for all ROIs (e.g. 64), no matter the range of intensities within a given ROI. Thus, the bin size is dependent on the range of intensities in a given ROI, i.e., not the same for all patients of a cohort. For FBS, the size of the bins gets fixed for all ROIs (e.g. 25 HU), no matter the range of intensities within a given ROI. Thus, the effective number of bins is dependent on the range of intensities in a given ROI and differs between patients in a cohort. Unlike FBN, FBS does conserve the relationship between image intensity and physical meaning, provided the minimum value of the first bin is fixed for all ROIs, as in this study. Finally, histogram equalization of the ROI imaging intensities can be performed before any discretization (FBN or FBS) is applied, leading to two other algorithms as defined by Vallières et al. [9]: FBNequal and FBSequal.

2.5. Description of comparative analysis

All features were calculated five times: for the chosen HU window, by shifting the window to the left (-10 HU, -20 HU), and by shifting the window to the right ($+10$ HU, $+20$ HU). For each feature vector (corresponding to the 95-patient cohort) and its four variants (± 10 HU, ± 20 HU), the four Spearman correlation coefficients were calculated (ρ_{10+} , ρ_{10-} , ρ_{20+} , ρ_{20-}). We stress that this is an outcome-independent analysis, and thus feature-outcome correlations are not involved. Then, ρ_{10+} , ρ_{10-} were averaged (denoted as ρ_{10}), and the ρ_{20+} , ρ_{20-} were averaged (denoted as ρ_{20}). We looked at median and lower bound of ρ_{10} and median and lower bound of ρ_{20} (for a certain feature set) as a function of the 21 energies. The lower bound was defined as the value above which 95% of the distribution lies. This was done for all radiomic features combined, and for the six texture feature types separately: grey level co-

occurrence matrix (GLCM, 25 features), grey level run length matrix (GLRLM, 16 features), grey level size zone matrix (GLSZM, 16 features), grey level distance zone matrix (GLDZM, 16 features), neighborhood grey tone difference matrix (NGTDM, 5 features), and neighboring grey level dependence matrix (NGLDM, 17 features). This allowed us to evaluate if certain texture types are more robust to changes in the HU window, and if this robustness depends on energy. This analysis was repeated for the four binning algorithms. To gain a sense of whether the positive and negative perturbations created a similar effect (which would justify our averaging approach), we compared median ρ_{10+} and ρ_{10-} as a function of the 21 energies. This was done for the six texture feature types separately (each texture type divided into FBN, FBNequal, FBS, and FBSequal algorithms). The same analysis was also performed with ρ_{20+} and ρ_{20-} .

To allow other researchers to verify our findings for this central analysis method, we repeated this comparative study on a publicly available dataset from The Cancer Imaging Archive for PET-CT imaging data of 298 patients from four different institutions in Quebec with histologically proven head-and-neck cancer [9–11]. The reason for not using the higher-quality radiotherapy-planning CT data was that during image acquisition, contrast enhancement was used for some patients but not others. To get a consistent CT scan for the entire cohort, we used the CT of PET-CT. We chose to use a head-and-neck dataset because the optimum HU range used in the paper should be applicable, which would not be the case for other anatomical sites such as lung. We have included the results of this study in the [Supplementary Material \(Tables A1–A4\)](#). To replicate this study, researchers can extract radiomic features using either our publicly available radiomic feature calculation code on GitHub (Section 2.2) or any other feature calculation software that offers the same functionality and verify our findings.

2.6. Explanation for not using concordance correlation coefficient

We do not use Lin's concordance correlation coefficient (CCC) [12]. If we were using the value of a radiomic feature directly, we would need to use CCC. But we use features for machine learning, where it is common practice to standardize a feature (set mean to 0 and standard deviation to 1). As Lin notes, CCC differs from Pearson correlation because of scale or location shifts; standardization fixes both these issues [13]. As the value of CCC is always less than or equal to the value of Pearson correlation, using CCC is overly punitive for our study, where the feature value itself has no biological meaning. Spearman (also known as rank) correlation quantifies the monotonicity of an association and behaves very similarly to Pearson for approximately linear relationships, with the benefit of being less influenced by outliers (common in medical data).

A perfect CCC of 1 is possible between two sets of observations X and Y only when a straight line through the origin fits the data, i.e., when $X = Y$. When the resegmentation range is changed, we know based on the mathematical definitions that for a given image, every texture matrix (GLCM, GLRLM, etc.) changes. Thus, features derived from these matrices should also change. Hence there is no reason to expect that a straight line fit through the origin is appropriate. In contrast, Spearman correlation of 1 is possible when the rank order of observations in X is the same as the rank order of observations in Y. For a machine learning algorithm trained on X to generalize well when applied to observations in Y, a high Spearman correlation is necessary, whereas a high CCC is not.

In the [Supplementary Material \(Figs. A8–A11\)](#), we demonstrate the effect of changing the resegmentation range by taking a publicly available non-medical images, adding or subtracting a fixed value from every pixel in the images (which is equivalent to shifting the resegmentation range to the left or right, respectively), and calculating the GLCM matrix for each image using a publicly available software (MATLAB). As expected, the GLCM matrix of an image is altered when the pixel values are

shifted uniformly.

2.7. Assessing effect of resegmentation range on predictive ability

There is also a need to understand how feature instability can affect *predictive ability* of a feature, which in this work we define as the Spearman correlation between the feature and the outcome (lymphadenopathy). This was studied in two ways. The first way was to record the average Spearman correlation of a feature and the outcome (i.e., the average predictive ability) for the top 10 texture features (out of 94 texture features) as a function of the 21 energies, top 10 meaning the ten texture features with the highest predictive ability. Since each texture feature has 64 ways of being extracted, the version with the highest predictive ability was chosen. This was done for the original feature, and the four HU window variants. The results for ± 10 HU were averaged, as were the results for ± 20 HU. The second way was to compare the overlap between the top 10 features based on the original HU range, and the top 10 features when using the four HU range variants, again as a function of the 21 energies. By overlap, we mean the fraction of top 10 features as derived using the original HU window that were still in the top 10 when a variant HU window was used. The results for ± 10 HU were averaged, as were the results for ± 20 HU.

3. Results

The empirically derived re-segmentation range was -50 to 350 HU ([Supplementary Material Fig. A1](#)). In [Fig. 1](#), we see the median and lower bound of ρ_{10} and ρ_{20} for all features as a function of the CT energy, separated based on the four binning algorithms. A tremendous dependence on algorithm choice is observed. For FBN and FBNequal algorithms, the correlation values are consistently above 0.9, indicating good stability. The peak is observed around 60–65 keV. By contrast, for FBS and FBSequal algorithms, while the median values are above 0.9, the lower bound falls as a function of energy, with poor performance over the entire spectrum (even reaching negative values for the FBS algorithm at the higher energies). FBNequal is the most stable algorithm, and FBS the least. The [Supplementary Material \(Figs. A2–A5\)](#) includes the results for each binning algorithm when separated by texture feature type (e.g., GLCM features, GLRLM features). To gain a deeper understanding of the instability observed in the FBS and FBSequal algorithms, we decided to study how this depends on the four bin sizes used (6.25, 12.5, 25, and 50 HU). These results are in the [Supplementary Material \(Figs. A7–A8\)](#). There is a steady worsening of stability as the bin size increases.

In [Fig. 2](#), we see the median of ρ_{10+} and ρ_{10-} separated into the six texture feature types (each texture type further split based on the four binning algorithms) as a function of the CT energy. Regardless of texture type and algorithm type, the ratio does not deviate from 1 by more than 1%, indicating very good symmetry. For all texture types, the most symmetric algorithm is FBNequal. The analogous plot for ρ_{20+} and ρ_{20-} can be found in the [Supplementary Material \(Fig. A6\)](#).

[Fig. 3](#) (left panel) shows the average Spearman correlation of a feature and the outcome (lymphadenopathy) for the top 10 texture features as a function of the 21 energies. Regardless of the discretization algorithm, the HU window variant features do have about the same predictive ability as the features obtained using the original HU window. For the FBN algorithm, there may be an increase in the mean correlation with energy, but given the size of the increase (from a correlation of about 0.35 at 50 keV to a correlation of 0.4 at 140 keV) and size of our dataset, we do not want to speculate whether this is a systematic and statistically significant increase. [Fig. 3](#) (right panel) shows, as a function of CT energy, the fractional overlap between the top 10 features based on the original HU range and the top 10 features when using the four HU range variants. We see that the less stable algorithms have less overlap, meaning that although the altered HU windows do not eliminate the existence of strong predictive features, the composition of the top 10

Fig. 1. Median (dashed line) and lower bound (solid line) of ρ_{10} (blue) and ρ_{20} (red) for all features as a function of the CT energy, separated based on the four binning algorithms.

Fig. 2. Median of the ratio between ρ_{10+} and ρ_{10-} separated into the six texture feature types (each texture type further split based on the four binning algorithms) as a function of the CT energy.

feature list changes by nearly half for the less stable algorithms.

4. Discussion

We believe this is the first systematic study of the impact of CT HU range used during the intensity discretization step of radiomic feature extraction. Although this is a fundamental part of handcrafted radiomic feature extraction, its importance has so far been overlooked. In many cases, the range used appears to be deferred to default settings without

clear justification and in some studies, the range is not even specified. Given the size of the effect for a change as small as ± 10 HU, particularly for the FBS and FBSequal algorithms, we recommend that future research projects pay greater attention to the CT HU range used, at least report the chosen range, and when needed, perform an analysis similar to what was presented here to determine the HU range that is suitable for their dataset. An exemplary publication should repeat its entire analysis by altering the selected HU range by ± 10 HU and explicitly quantify the effect of such a change so as to convince the reader of the

Fig. 3. Left panel: Mean Spearman correlation between feature and outcome for top 10 texture features separated into the four binning algorithms as a function of the CT energy. The black line represents the features calculated using the original HU window (blue line = ± 10 HU, red line = ± 20 HU). Right panel: Fractional overlap, as a function of the CT energy, between the top 10 features based on the original HU range and the top 10 features when using the HU range variants, separated into the four binning algorithms. Blue line = overlap between ± 10 HU and default, red line = between ± 20 HU and default.

robustness of the findings. We note that some high-impact publications have used the FBS algorithm and are potentially susceptible to this instability [14,15]. The HU range established by our empirical algorithm is consistent with the experience of our expert radiologist.

As mentioned previously, there were two main aims of this study: (1) to evaluate if feature robustness depended on the discretization algorithm, and (2) to investigate if this robustness depended on CT electron energy. Each figure in this paper addressed both aims. Fig. 1, with four subfigures, showed the importance of the discretization algorithm (FBS and FBSequal were unstable, whereas FBN and FBNequal were not), and each subfigure showed the importance of the different virtual energies (the instability for FBS and FBSequal worsened with increasing virtual energy). Fig. 2, with 4 lines per subfigure, showed the importance of the discretization algorithm (while all algorithms showed very good symmetry, FBNequal was the most symmetric), and each subfigure investigated the importance of the different virtual energies (there were no energy-dependent trends). Fig. 3, with four subfigures per panel, showed that discretization algorithm did not affect predictive ability but did affect fractional overlap (less stable algorithms had less overlap), and none of the subfigures showed significant energy-dependent trends.

As a field matures, it becomes increasingly important to pay attention to all relevant sources of uncertainty, not just the obvious sources such as differences between CT machines from different manufacturers or inter-observer variability in terms of ROI delineation. In the field of radiation therapy simulation, Monte Carlo calculations of radiation dose depends on many computational parameters, the effects of which may introduce sub-percent uncertainty. Publications in this field are subject to practice guidelines such as TG-268 [16] which state the mandatory reporting requirements for simulation parameters. We believe similar practice guidelines are essential for widespread clinical translation of radiomic biomarkers. If a standard radiomic feature extraction package is used without evaluating the suitability of default settings for the anatomical site or lesion of interest, the analysis results are likely to be sub-optimal. We foresee the eventual establishment of experimentally validated HU range(s) for different anatomic sites and pathologies following our methodology.

We emphasize that we do not mean to imply that radiomic feature extraction should only be limited to one set of extraction parameters. We believe there is a benefit in extracting radiomic features using different parameters, as this might help us capture different biologically relevant phenotypes [17]. This is why we calculate each texture feature 64 ways by using all combinations of the following hyper-parameters: 4 voxel

sizes, 4 binning algorithms, and 4 gray level discretizations. Our central point is that one such extraction parameter, i.e., CT HU range, must be optimally chosen for the anatomical site in question, and then clearly reported in any publication to ensure replicability of the findings. We do not believe that the ± 10 or ± 20 HU variations used in our stability studies are capturing new biological information; they simply allow us to quantify the systematic uncertainty arising from the choice of the CT HU range.

In addition to highlighting the general importance of HU range and the binning algorithm used during feature extraction, this study reveals variations based on the virtual monochromatic image energy used for feature calculation. There are increasing studies demonstrating that multi-energy datasets from dual energy CT scans can be used to improve prediction performance in radiomics [8]. Understanding the stability of radiomic features at different energies and the optimal binning algorithm will be important for future radiomic applications using multi-energy CT and their generalizability.

The main limitation of this study is that the results are based on one anatomic area and pathology (HNSCC). However, the purpose of this study was not to determine the specific impact of HU range and binning algorithm for every anatomic site or pathology, a task that would not be possible in a single study. Rather, this work highlights the importance and potential impact of these parameters and approaches, necessitating additional investigations and optimization in future radiomic research. We do not intend for other researchers to assume that FBS and FBSequal will always be the least stable algorithms regardless of the choice of dataset. For example, for the analysis we performed on the public dataset, while FBS was indeed the least stable algorithm, FBSequal was quite stable. The stability of the discretization algorithm will in fact depend on many factors such as the imaging contrast within the region of interest analyzed and the specific types of texture features calculated, as shown in [Supplementary Material](#). For this reason, we encourage the community to always assess and take into account the impact of the HU window on their specific radiomic study. Another limitation of this study was that we did not focus on different clinical endpoints. However, we demonstrated significant fluctuations in features associated with one clinical endpoint, cervical lymphadenopathy, depending on the binning algorithm used. This illustrates the importance of such an analysis for the derivation and use of radiomic biomarkers.

While our dataset of 95 patients may seem small, we want to emphasize that this is a methodological paper meant to highlight the importance of HU range. We do not expect the reader to use quantitative

results of this paper directly, but rather to repeat the study for their own dataset. The sample size is adequate for demonstrating our methodology.

5. Conclusion

We presented a detailed analysis of how the choice of HU window used to discretize CT data affects radiomic features. On our dataset, we found that FBN and FBNequal algorithms showed good stability for changes of ± 10 HU and ± 20 HU. By contrast, FBS and FBSequal algorithms showed poor stability. The FBNequal algorithm was the most symmetric under perturbation. We suggest that future analyses take this source of uncertainty into account when evaluating the robustness of their radiomic signature in order to increase the likelihood of replicability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

AC acknowledges partial support by the CREATE Medical Physics Research Training Network grant of the Natural Sciences and Engineering Research Council (Grant No.: 432290) and the Canadian Institutes of Health Research Foundation Grant FDN-143257. MV acknowledges funding from the Canada CIFAR AI Chairs Program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.ejmp.2021.07.023>.

References

- [1] Gillies R, Kinahan P, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* 2016;278(2):563–77.

- [2] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012;48(4):441–6.
- [3] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging* 2012;30(9):1234–48.
- [4] Larue R, Defraene R, De Ruyscher D, Lambin P, Van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol* 2017;90(1070):20160665.
- [5] Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Bist V, et al. An empirical approach for avoiding false discoveries when applying high-dimensional radiomics to small datasets. *IEEE TRPMS* 2018;3(2):201–9.
- [6] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020;295(2):328–38.
- [7] Forghani R, Savadjiev P, Chatterjee A, Muthukrishnan N, Reinhold C, Forghani B. Radiomics and artificial intelligence for biomarker and prediction model development in oncology. *Comput Struct Biotechnol J* 2019;17:995.
- [8] Forghani R, Chatterjee A, Reinhold C, Pérez-Lara A, Romero-Sanchez G, Ueno Y, et al. Head and neck squamous cell carcinoma: prediction of cervical lymph node metastasis by dual-energy CT texture analysis with machine learning. *Eur Radiol* 2019;29(11):6172–81.
- [9] Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep* 2017;7(1):1–4.
- [10] Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJ, et al. Data from Head-Neck-PET-CT. *Cancer Imaging Arch* 2017. <https://doi.org/10.7937/K9/TCIA.2017.8oje5q00>.
- [11] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013;26(6):1045–57.
- [12] Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;255–68.
- [13] Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Saif S, et al. Creating robust predictive radiomic models for data from independent institutions using normalization. *IEEE TRPMS* 2019;3(2):210–5.
- [14] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Com* 2014;5(1):1–9.
- [15] van Dijk LV, Brouwer CL, van der Schaaf A, Burgerhof JG, Beukinga RJ, Langendijk JA, et al. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. *Radiother Oncol* 2017;122(2):185–91.
- [16] Sechopoulos I, Rogers DW, Bazalova-Carter M, Bolch WE, Heath EC, McNitt-Gray MF, et al. RECORDS: improved reporting of monte Carlo Radiation transport studies: report of the AAPM Research Committee Task Group 268. *Med Phys* 2018;45(1):e1–5.
- [17] Vallières M, Freeman CR, Skamene SR, El Naqa IA. Radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol* 2015;60(14):5471..