Effect of Reconstruction Parameters on the Quantitative Analysis of Chest Computed Tomography

Journal of Thoracic Imaging

Effect of reconstruction parameters on the quantitative analysis of chest computed tomography

--Manuscript Draft--

Manuscript Number:	JTI-18-110R1
Full Title:	Effect of reconstruction parameters on the quantitative analysis of chest computed tomography
Article Type:	Review Article (Invited)
Section/Category:	Pulmonary/thoracic
Keywords:	Computed Tomography; reconstruction; reconstruction; Radiomics; Radiomics; emphysema; emphysema; volumetry; volumetry; quantification; quantification
Corresponding Author:	Jin Mo Goo, MD Seoul National University Hospital Seoul, KOREA, REPUBLIC OF
Corresponding Author's Institution:	Seoul National University Hospital
Order of Authors:	Hyungjin Kim
	Jin Mo Goo, MD
	Yoshiharu Ohno
	Hans-Ulrich Kauczor
	Eric Hoffman
	James Gee
	Edwin van Beek
Manuscript Region of Origin:	KOREA, REPUBLIC OF
Abstract:	Summary Quantitative features obtained from CT scans are being explored for clinical applications. Various classes of quantitative features exist for chest CT including radiomics features, emphysema measurements, lung nodule volumetric measurements, dual energy quantification, and perfusion parameters. A number of research articles have shown promise in diagnosis and prognosis prediction of oncologic patients or those with diffuse lung diseases using these feature classes. Nevertheless, a prerequisite for the quantification is the evaluation of variation in measurements in terms of repeatability and reproducibility, which are distinct aspects of precision but are often not separable from each other. There are well-known sources of measurement variability including patient factors, CT acquisition (scan and reconstruction) factors, and radiologist (or measurement-related) factors. The purpose of this article is to review the effects of CT reconstruction parameters on the quantitative imaging features and efforts to correct or neutralize variations induced by those parameters.

August 23, 2018

Dr. Mark L. Schiebler

Guest Editor, Journal of Thoracic Imaging

Dear Dr. Schiebler,

Thank you very much for the review and comments on our paper. Herein, we submit our revised manuscript entitled "Effect of reconstruction parameters on the quantitative analysis of chest computed tomography" (JTI-18-110).

We made efforts to follow the reviewers' valuable recommendations and suggestions as best as we could and we addressed all the reviewers' concerns in a point by point manner in the response letter and in the manuscript. We added general description regarding the radiomics analysis with an exemplary figure of a work flow. We also expanded other graphical contents as recommended by the reviewer 2. As for the discussion around the merits of CT quantification compared with visual assessment, we supposed that this issue was beyond the scope of our review article, which specifically focused on the effects of reconstruction parameters on the CT quantitative features.

We hope that our response and revision can alleviate the reviewers' concerns. Thank you once again and we look forward to hearing good news.

Sincerely Yours,

Jin Mo Goo, MD, PhD

Department of Radiology, Seoul National University College of Medicine, 101, Daehak-ro, Jongno-gu, Seoul, 03080, Korea. Tel: 82-2-2072-2624, Fax: 82-2-743-6385. E-mail: jmgoo@plaza.snu.ac.kr

Point-by-point response letter to the reviewers' comments

Reviewer #1: Very well written and organized manuscript.

Conclusions make sense.

Suggestions:

1. Page 8 second §: "make" is unclear. Please clarify.

→ Thank you for your review and comments. We changed the word as 'manufacturer' in the manuscript (page 8).

2. Page 8 first §, line4: please give tumor size range.

→ The authors of the study did not report the tumor size range. We described as such in the manuscript (page 8).

Reviewer #2: Congratulations on the nice review

Comments:

1. Well organized but a more meaningful discussion around radiomics in general would be helpful. How does it work? How will it help with pulmonary quantification? A figure?

→ We appreciate your comments. We additionally described regarding the radiomics features and the general process of radiomics analysis in page 5. We added figure 1 as an example of a radiomics analysis work flow as you suggested.

2. The figures are limited and need expansion

→ We agree with your opinion and we prepared additional figures with respect to the effect of reconstruction kernel on the radiomics features (Figure 3; page 6 and 31) and the effect of noise reducing strength of iterative reconstruction algorithms on the emphysema quantification (Figure 4; page 10 and 32).

3. A concluding paragraph on where things are going would be helpful

→ Please refer to page 18 for the future directions of reproducibility analysis. We emphasized three points: 1) standardization of methodology and development of standardized CT phantoms and software programs; 2) necessity for comprehensive evaluation of several quantitative feature classes; and 3) task-based analysis to reveal the clinical relevance of measurement variability.

4. Discussion around the merits of quantification vs visual assessment as marker or measures of treatment response would be helpful.

→ Thank you for the suggestion. However, we assume that the discussion regarding the merits and demerits of quantification compared with the visual assessment is beyond the scope of our article. In this review article, we specifically focused on the effect of several reconstruction parameters on the CT quantitative features. Thus, the pros and cons of CT quantification should be investigated and summarized as a separate review.

Effect of reconstruction parameters on the quantitative analysis of chest computed tomography

Running head: Effect of CT reconstruction on quantification

Hyungjin Kim, MD,[1] Jin Mo Goo, MD, PhD,[1, 2] Yoshiharu Ohno, MD, PhD, [3] Hans-Ulrich Kauczor, MD,[4] Eric A. Hoffman, PhD, [5] James C. Gee, PhD, [6] Edwin J. R. van Beek, MD, PhD, [7]

[1]Department of Radiology, Seoul National University College of Medicine, and Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Korea; [2]Cancer Research Institute, Seoul National University, Seoul, Korea; [3]Division of Functional and Diagnostic Imaging Research, Department of Radiology, Kobe University Graduate School of Medicine, Kobe, Japan; [4]Department of Diagnostic and Interventional Radiology, University of Heidelberg, Translational Lung Research Center Heidelberg, Heidelberg, Germany; [5]Departments of Radiology, Medicine and Biomedical Engineering, University of Iowa, Iowa City; Iowa, USA; [6]Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA; [7]Edinburgh Imaging, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK

Conflicts of Interest and Source of Funding: Yoshiharu Ohno received a research grant from Canon Medical Systems. Edwin J. R. van Beek is a member of advisory boards of Imbio and Aidence. He is also the owner/founder of Quantitative Clinical Trials Imaging Services. The

remaining authors declare no conflicts of interest.

Contact Information for Corresponding Author:

Jin Mo Goo, MD, PhD. Department of Radiology, Seoul National University College of Medicine, 101, Daehak-ro, Jongno-gu, Seoul, 03080, Korea. Tel: 82-2-2072-2624, Fax: 82-2-743-6385. E-mail: jmgoo@plaza.snu.ac.kr

Effect of reconstruction parameters on the quantitative analysis of chest computed tomography

Summary

Quantitative features obtained from CT scans are being explored for clinical applications. Various classes of quantitative features exist for chest CT including radiomics features, emphysema measurements, lung nodule volumetric measurements, dual energy quantification, and perfusion parameters. A number of research articles have shown promise in diagnosis and prognosis prediction of oncologic patients or those with diffuse lung diseases using these feature classes. Nevertheless, a prerequisite for the quantification is the evaluation of variation in measurements in terms of repeatability and reproducibility, which are distinct aspects of precision but are often not separable from each other. There are well-known sources of measurement variability including patient factors, CT acquisition (scan and reconstruction) factors, and radiologist (or measurement-related) factors. The purpose of this article is to review the effects of CT reconstruction parameters on the quantitative imaging features and efforts to correct or neutralize variations induced by those parameters.

Key Words

computed tomography; reconstruction; radiomics; emphysema; volumetry; quantification

Introduction

Quantitative computed tomography (CT) features vary from simple uni-dimensional measurements to those calculated from complex three-dimensional (3D) matrices of pixel value distribution. Although deep learning is beyond the scope of this article, it is also a way of automatically extracting and learning a number of quantitative readouts from simple to complex features. Ideally, a quantitative imaging readout can be used as an imaging biomarker, which is defined as an objectively measured characteristic derived from an in vivo image as an indicator of normal biologic processes, pathogenic processes, or response to a therapeutic intervention.[1]

Before the clinical application of any quantitative features, linearity, bias, and precision have to be scrutinized.[2] Evaluation of these statistical and metrologic methodologies is essential for the translation of research-level quantification to clinical practice. According to Sullivan et al.,[2] linearity is the ability to provide measured values that are directly proportional to the true values. Bias, commonly termed as accuracy, is the difference between the mean of measurements determined from the same object and true value.[2] Precision is variability of the measurement, which can be classified into repeatability and reproducibility.[2] Repeatability refers to the variation of feature values at the repeated identical imaging condition, while reproducibility is the variation of the measurement according to the different conditions (i.e., CT scanners, protocols, and institutions).[2]

Reproducibility of measurements has been studied extensively in the field of radiology as the range of variation is crucial to determine the true biological change in vivo and to determine the normal range of observations. There are a lot of potential sources of variations including imaging acquisition factors (technician, scanner, manufacturer, acquisition

parameters, and reconstruction parameters) and radiologist factors (operator and software tool). Any of these factors can affect the reproducibility of measurements.

The importance of reproducibility especially at a longitudinal study is that measurement change can be attributable to either true change or measurement variation. If the measurement variation is large, the clinical decision cannot be made with confidence. In addition, if inter-individual dynamic range (measurement differences between individuals) of a certain feature is smaller than the potential measurement variation, then that feature may not be utilized as an imaging biomarker. Nevertheless, conversely, it means that recognition of imaging factors that generate variation raises an issue to standardize quantification.

In this review, we aimed to describe implications of several CT reconstruction parameters on the quantitative imaging features from the aspect of measurement reproducibility. We dealt with radiomics features, emphysema quantification and lung nodule volumetry. Influence of reconstruction parameters such as iterative reconstruction algorithms, reconstruction kernels, and slice thickness will be discussed.

Radiomics features

Radiomics analysis refers to high-throughput extraction of quantitative features including intensity or texture features from images. It is usually performed in a following order: lesion segmentation, feature extraction, feature selection, model training and validation (Fig. 1). Radiomics analysis enables to capture and calculate independent imaging features (e.g. tumor heterogeneity) which may or may not be visible to the human eyes. Radiomics features include first-order features which do not consider spatial relationships among voxel values and second-order features from gray-level co-occurrence matrix (GLCM), gray-level run length matrix, gray-level size zone matrix, and neighborhood gray-tone difference matrix (NGTDM). ~~I~~This method has potential to further promote the role of imaging in the era of precision medicine.[3] A number of studies based on the radiomics analysis for diagnosis, cancer staging, prognosis prediction, treatment response monitoring, and surveillance have been reported to date.[3] Nevertheless, standardization of analysis protocols, reproducibility of features, and redundancy of extracted information are the remaining concerns for the radiomics approach.

Commented [C1]: Reviewer 2-1

Effect of iterative reconstruction algorithm

A few studies have shown the effect of iterative reconstruction (IR) on the radiomics features (Fig. ~~1~~2).[4-6] Kim et al.[4] reported that most of the first-order intensity features and second-order ~~gray-level co-occurrence matrix (GLCM)~~GLCM-based features showed significant differences between filtered back projection (FBP) and Sinogram Affirmed Iterative Reconstruction (SAFIRE) using chest CT scans of patients with pulmonary tumors. Interestingly, features were also substantially influenced by the noise reduction strength of SAFIRE (level 3 vs. 5). Size features, entropy, and GLCM entropy were the most robust features (coefficient of variation ≤5%) when inter-reader variability in tumor segmentation and

inter-reconstruction algorithm variability were taken into account. Solomon et al.[5] also revealed that five first-order features and two GLCM-based features were affected by model-based iterative reconstruction (MBIR). These studies showed that the feature values differed significantly according to the reconstruction algorithm of CT scans and the range of variation could be substantial. However, none of those analyzed the potentially substantive clinical implication of the measurement variability induced by IR in terms of diagnosis or prognostication. Measurement variation of features may affect the actual performance of a diagnostic or prognostic model by erroneously increasing the overlap between the classes of label data. This topic warrants future investigations.

Effect of reconstruction kernel and slice thickness

As for the effect of reconstruction kernel (Fig. 3) and slice thickness on the radiomics features, Lu et al.[7] analyzed concordance correlation coefficients (CCCs) of the radiomics features between reconstruction settings of different slice thickness (1.25, 2.5, and 5 mm) and kernels (lung and standard kernel). They found that the agreement levels of changing reconstruction kernels were lower than those of changing slice thickness. Obviously, changing both the reconstruction kernel and the slice thickness (slice thickness of 1.25 mm and lung kernel vs. slice thickness of 5mm and standard kernel) resulted in the worst agreement (CCCs <0.51). Among multiple radiomics features, size, mean density, coarse boundary morphology, and coarse texture features were relatively robust to the changes in slice thickness and kernels. Intriguingly, boundary sharpness and fine texture features, which contained detailed morphological or textural information, were vulnerable to the reconstruction settings. Zhao et al.[8] performed a similar study using same-day repeat CT scans. They revealed that the radiomics features obtained from thin-section (1.25 mm) images had higher inter-scan

Commented [C2]: Reviewer 2-2

agreement. Laplacian of Gaussian features obtained from the standard kernel were more reproducible than those from the lung kernel. Zheng et al.[9] performed a phantom study with 3D printed textured lesions. They concluded that among multiple variables of radiation dosage (0.67, 1.42, and 5.80 mGy), reconstruction algorithm (FBP and IR), kernel (standard, soft, and edge), and slice thickness (0.6 and 5 mm), the last two parameters were most influential. In addition, features from thin-slice and edge reconstruction kernel were more accurate and reproducible. The discrepancy of the study results between Zhao et al.[8] and Zheng et al.[9] (standard vs. edge kernel for the reproducibility) might be due to differences in extracted radiomics features. Some features might be more vulnerable to the sharp kernel, while others might show higher variation to the standard kernel.

A task-based assessment for the implication of the reconstruction parameters was performed by He et al.[10] Their investigation demonstrated a link between a variation in feature values and the actual diagnosis. They conducted regression analysis with the radiomics features obtained from several different CT scans (enhanced vs. non-enhanced; 1.25 vs. 5 mm; standard vs. lung kernel) for the differentiation of malignant lung nodules. They concluded that the optimal combination for the highest diagnostic performance was non-contrast, thin-slice, and standard kernel. The combination of enhancement status and reconstruction parameters had substantial effects on the feature selection and subsequently on the performance of diagnostic models.

There have been attempts to reduce the variation of radiomics features. Larue et al.[11] focused on the resampling method to reduce the variability or dependency of radiomics features on the slice thickness. They found that most radiomics features were affected by the slice thickness and this variation could be reduced by resampling of voxel sizes using cubic or linear

interpolation before feature extraction. They also demonstrated that linear interpolation (voxel size resampling into 1x1x3 mm^3) resulted in feature stability in 48% of the radiomics features. For the kernel-induced variability, a recent study[12] suggested that 3D noise power spectrum peak frequency and region of interest maximum intensity could be used as correction factors.

Effect of pixel size

Given the fact that slice thickness causes variation in the radiomics features, it is easily understandable that the effect of pixel size or in-plane resolution on the radiomics features would also be considerable. Mackin et al.[13] compared the intra-patient variability caused by the variation in pixel size using CT scans reconstructed with 5 different pixel sizes ranging from 0.59 to 0.98 mm. They retrospectively reconstructed CT scans of 8 lung cancer patients (tumor size not reported) with varying field-of-views from 30 to 50 cm in 5 cm increments. The intra-patient variability (overall CCC) was larger than the inter-patient variability in 79% of the features, which implied that the variation caused by the pixel size may obscure true inter-individual differences. For the correction of variation in pixel sizes, they combined image resampling (bilinear interpolation to 1 mm/pixel) with Butterworth low-pass filtering in the frequency domain. After correction, the intra-patient variability was relatively large in only 10% of the features. Shafiq-ul-Hassan et al.[14] reported a similar finding. Among 213 features, variation in 42 features diminished significantly after resampling (percent coefficient of variation <30). In their study, 150 features were reproducible irrespective of the voxel sizes and 21 had substantial variation before and after voxel size resampling.

Commented [C3]: Reviewer 1-2

Effect of CT scanner

CT scanner ~~make~~ manufacturer and model is another cause of variability for the radiomics features. Mackin et al.[15] performed 17 CT scans using CT scanners from four

Commented [C4]: Reviewer 1-1

manufacturers. Although the scanning parameters were not equal across multiple protocols and CT scanners, the interesting finding was that the radiomics feature values obtained from the same vendor were grouped together at hierarchical clustering. For instance, entropy or strength values obtained from Philips scanners showed negative deviation from the normalized mean of 0 and those from Siemens and Toshiba (currently, Canon Medical Systems) scanners exhibited positive deviation from the normalized mean. Variation due to CT scanners was observed even after acquiring CT scans with near-identical scanning parameters.[16] Mahmood et al.[16] obtained ~~neighborhood gray-tone difference matrix (~~NGTDM~~)~~ and GLCM features from three CT scanners. Scanning parameters including voxel size, radiation dosage, pitch, and kernels were matched across the CT scans. However, NGTDM and GLCM features were not reproducible among different scanners with CCCs of less than 0.9. Scanner-induced variability was also reported for the image filtration-based features.[17] Scanner-induced variation is probably due to detector design, beam spectra, calibration methods, and quality control/maintenance of CT machines.

A robust correction factor or development of a less variable feature, which is clinically relevant at the same time, would be required. From this aspect, Chen-Mayer et al.[18] proposed a 5-step calibration procedure using a single parameter to describe scanner dependent contribution to the pixel value. They mapped CT numbers to 80 keV and demonstrated that reproducibility of the pixel values markedly improved after calibration (standard deviation less than 1 Hounsfield Unit [HU]). As radiomic features were not evaluated in this study, further investigation using the described calibration procedure is required.

Emphysema quantification

CT provides valuable quantitative parameters for the prediction of the degree of airflow obstruction at pulmonary function testing and the risk of chronic obstructive pulmonary disease (COPD) exacerbations.[19] CT quantification for COPD evaluation can be categorized into emphysema quantification such as percentage low-attenuation area (%LAA; sometimes termed as relative area -950 HU or emphysema index), airway measurement (airway wall thickness, wall area, lumen diameter, or internal perimeter) and air trapping analysis. These quantitative features have been investigated in terms of reproducibility.

Effect of iterative reconstruction algorithm

Studies to date have consistently shown that IR has significant influence on the quantification of LAA. A number of commercial IR algorithms have been assessed including SAFIRE from Siemens Healthcare;[20, 21] adaptive statistical iterative reconstruction (ASIR) and MBIR from GE Healthcare;[21, 22] iDose from Philips Healthcare;[23] and adaptive iterative dose reduction using 3D processing (AIDR3D) from Canon Medical Systems.[24, 25] These referenced studies demonstrated that LAA would be underestimated if IR is applied. IR changes the distribution of pixel values in the extremes of the attenuation histogram.[20] That is, the mean of pixel value is preserved, but the standard deviation is reduced and the density histogram becomes more sharply peaked.[26] Such convergence toward the mean leads to decrease in the number of pixels below a certain cutoff such as -950 HU. In addition, such phenomenon is accentuated when the IR strength or noise reduction level increase.[20, 21] This means that the grade of denoising affects LAA measurement. Underestimation of LAA was greater with MBIR than with ASIR (Fig. 4).[21, 22]

Commented [C5]: Reviewer 2-2

A beneficial effect of IR in emphysema quantification was reported by Japanese researchers.[24, 25] They suggested that the agreement of LAA between standard-dose and low-dose (or ultra-low-dose) CT scans could be improved by using IR.[24, 25] That is, increased image noise in low-dose CT, which may result in an increase in LAA (overestimation), could be mitigated by the application of IR. This finding is promising given that the effect of noise on quantification, such as in cases of low-dose setting or scanning large patients, can be relieved by IR. Another potential strength of IR was proposed by Choo et al.[22] They performed a phantom study using a CTP674 Lung Phantom which included polycarbonate tubes simulating human airways. They compared the accuracy of airway measurements between the reconstruction algorithms (FBP, ASIR, and MBIR). Notably, MBIR exhibited the most accurate results for the measurements of luminal area, wall area, and wall thickness.

A more challenging method for the preservation of pixel value distribution of FBP after application of IR was suggested by Rodriguez et al.[26] They compared pixel value distribution of phantom inserts among CT scan combinations of different kernels (standard and sharp) and reconstruction algorithms (FBP, ASIR, and MBIR). In their study, combination of bone kernel with ASIR had similar pixel value distribution with the pair of standard kernel and FBP. It should be noted that various emphysema quantification parameters should be tested to guarantee the comparability of feature values between that combination and this approach should be analyzed for other CT scanners and vendors.

Effect of reconstruction kernel and slice thickness

It is now a well-known fact that the reconstruction kernel should be kept constant for the longitudinal emphysema quantification. A number of studies have demonstrated that application of a sharp kernel can cause an erroneous increase of LAA (Fig. ~~2~~5).[27-30] Sharp

kernel modifies HU of interface pixels between structures with significantly different attenuation coefficients.[31] Therefore, detection of tissue or material interface can be affected by the kernel used.[28] In addition, increased image noise associated with sharp kernel broadens the width of pixel value histogram, which subsequently results in increased LAA as abovementioned. The effect of slice thickness can be explained in the same way. Thin-slice reconstruction can also cause increment in LAA.[27, 30, 32, 33]

Madani et al.[33] reported correlations between LAA from smooth kernel-reconstructed CT scan and macroscopic and microscopic morphometric measurements. Hochhegger et al.[28] proposed that a standard kernel should be preferred given the larger longitudinal variability of LAA at sharp kernel CT scans. Gierada et al.[27] also presented the pros of smooth kernel based on the stronger correlation of emphysema index with histological alveolar wall distance measurements at smooth kernels.

To relieve the variability in emphysema quantification caused by the reconstruction kernel, Gallardo-Estrella et al.[34] suggested a normalization method of different reconstruction kernels by frequency band decomposition with hierarchical unsharp masking to standardize the energy in each band to a reference value. By using this method, calculated energy coefficients can be applied to various kernel images to create normalized images. This method effectively reduced variation in emphysema quantification caused by the kernel. Gallardo-Estrella et al.[35] also reported that emphysema quantification after pixel resampling to 3 mm slice thickness, normalization, and bullae analysis to minimize variability in slice thickness, kernel, and noise led to better prognostication of all-cause mortality and lung cancer mortality. Another group of researchers compared three kernel normalization methods and reported that edge-preserving frequency decomposition was the best normalization method for quantification of emphysema

index.[36] Ohkubo et al.[37] reported that the ratio of modulation transfer functions for the two different kernels could be used as a filter function for kernel conversion. However, emphysema quantification was not investigated in this study. More recently, deep learning was applied for the kernel conversion from sharp to smooth for the quantification of LAA.[38] Pairs of standard and sharp kernel images were fed to convolutional network for training and the deep learning model produced LAA of converted images (8.87 ± 6.20%, converted B50f; 27.65 ± 7.28%, original B50f) similar to that of the standard kernel images (10.82 ± 6.71%, B30f).[38]

Effect of quantification software

Another potential source of variation in emphysema measurement is the software program for automated quantification, which causes variation in lung segmentation, airway segmentation and subsequent quantification.[39] Wielputz et al.[39] compared three software tools including one in-house software and two commercial software products from major vendors. They revealed that the inter-software variability of emphysema index was higher than the measurable progression with median differences from -5.0 to -1.7%. Limits of agreement were as wide as -25.5 to 18.8%.[39] They also stated that such a large variability range could affect patient inclusion or exclusion for endobronchial valve treatment.[40] Therefore, the researchers described that the software cannot be used interchangeably for the longitudinal follow-up or post-treatment evaluation.[39] Inter-software variability for the lobe-based quantification was reported by Lim et al.[41] Four fully automated lobar segmentation tools were analyzed and limits of agreement for emphysema index was as large as -7 to 14%.[41] Patients with inhomogeneous emphysema distribution, who are suitable for surgical or bronchoscopic lung volume reduction surgery, showed higher inter-software variability due to greater distortion of normal anatomy.[41]

Lung nodule volumetry

Volumetric measurement has a relatively long history compared to other quantitative features, which goes back to the late 1990s. During the past decades, its technical as well as clinical performance have been investigated by researchers.[42] Volume measurement has several merits over diameter measurement. First, volume is more representative of the true dimension of a nodule considering substantial diameter variation within a nodule.[43] Second, it has potential to further stratify the intermediate risk category which is determined by diameter measurement.[43] Third, semi-automated volumetry may provide reproducible measurements.[44]

Effect of iterative reconstruction algorithm

To date, most studies have reported that the effect of IR on the lung nodule volumetry was not clinically relevant.[45-50] In detail, there was either no statistically significant difference in volumetric results between FBP and IR or the magnitude of IR-induced variation was smaller than the inter-reader or inter-scan measurement variability (Fig. ~~3~~6). Extensive studies have been performed, which investigated the effect of IR with and without other factors such as radiation dose on the lung nodule volumetry in terms of bias and precision. IR algorithms including ASIR and MBIR;[51, 52] SAFIRE;[48] AIDR 3D and forward projected model-based iterative reconstruction (FIRST; Canon Medical Systems]);[53, 54] iDose and ~~iterative model reconstruction (IMR; Philips Healthcare)~~IMR[45-47, 49, 50, 55] were investigated using phantoms or in vivo lung nodules. Considering that the inherent contrast between the pulmonary nodules and background parenchyma is high, it can be expected that the margin segmentation might not be much affected by the reconstruction algorithms. For the phantom nodules, researchers have shown that even the volumetric measurements of subsolid nodules, were robust to the change in the reconstruction algorithms.[46, 47] Regarding the subsolid nodules in vivo, Cohen et

al.[52] compared the volumetric parameters between FBP and MBIR and found that volume and mass of the subsolid nodules as a whole and their solid components were measured significantly larger when using MBIR.[52] Nevertheless, differences between the reconstruction algorithms (up to 6.6%) were within the range of intra- and inter-reader variability.[52] A few studies suggested that the measurement accuracy for the volumetric analysis of subsolid nodules was higher with IR at extremely low radiation dose levels (10-20 mA).[53-55]

Effect of reconstruction kernel

The effect of the reconstruction kernel on lung nodule volumetry is somewhat controversial. Ravanel et al.[56] reported that lung kernel yielded the least bias among seven different reconstruction kernels but they also stated that no single kernel consistently yielded both the least biased and the most precise volume estimations. Ko et al.[57] also reported that a high-frequency kernel provided less bias than a low-frequency algorithm. However, Wang et al.[58] performed three repeated measurements of solid lung nodules to analyze repeatability and found that soft kernel was better than sharp kernel, particularly for non-smooth-round nodules. For the artificial subsolid nodules, Scholten et al.[59] reported that absolute percentage error was not influenced by the kernel. More recently, Christe et al.[60] compared soft and sharp kernels for 113 lung nodules and reported that the mean volume was measured generally higher using the soft kernel. The relative difference of volume between the kernels ranged from 5.9% to 20.5% depending on the software programs.[60] Limits of agreement between the kernels for one of the tools ranged from -7.9% to 49%.[60]

Based on these studies, it is difficult to choose an optimal kernel setting for the lung nodule volumetry as both bias and precision are the essential components of measurements. However, it is obvious that consistency in reconstruction kernel is definitely required given the

substantial difference of volume measurement according to the kernel selection. Further research for the recent scanners and software programs based on both phantom and in vivo nodules are warranted for conclusive results.

Effect of slice thickness

Slice thickness particularly matters for the small lung nodules as the proportion of surface voxels increases in the small nodules.[42] If section thickness increases, the margin of nodules becomes blurred due to partial volume averaging of surface voxels.[42] Accordingly, the estimated volume would increase. Volume overestimation at thick-section CT scans was reported by Ravanel et al.[56] using synthetic nodules and Zhao et al.[61] using in vivo metastatic nodules. Other researchers reported similar findings.[62, 63] In addition, Wang et al.[58] demonstrated that slice thickness was associated with repeatability of measurements and the relative volume differences in 2 mm thickness (22.5%) were greater than those in 1mm thickness (8.9%). Voxel resampling as mentioned in the previous section would be helpful for the longitudinal follow-up of nodules on CT scans with different slice thicknesses. The error correction equation can also be applied.[64]

Effect of volumetry software programs

General consensus from the published articles is that the software program or its internal segmentation algorithms have a substantial effect on the measured nodule volume.[65-70] It is to be noted that the software-induced variation can sometimes be greater than the nodule volume growth cutoff (25%),[66, 67] which is currently used at the British Thoracic Society guideline[71] or Dutch-Belgian lung cancer screening trial (NELSON).[72] This indicates that the same software or segmentation algorithm should be consistently used for the surveillance of nodules in order not to misinterpret measurement variation as true growth. This would be

particularly important for the lung cancer screening setting or treatment response evaluation. Zhao et al.[67] performed volume doubling time (VDT) calculation using CT scans of three time points with three software tools. They reported that the consensus on VDT categorization at follow-up was less than 50% when the results from different software programs were compared.[67] In addition, Kalpathy-Cramer et al.[69] revealed that the most accurate segmentation algorithm was not the most repeatable. Therefore, care should be taken for the selection of a segmentation algorithm based on the thorough evaluation of both bias and precision. This is important as the screening-detected nodules are categorized by the baseline nodule volume and also classified by VDT at follow-up scans.[73]

Further directions of reproducibility analysis

Many studies for the quantitative imaging biomarkers have been conducted thus far. However, most of them are still research-based and most are not ready for actual clinical application. One of the main reasons is the standardization of methodology. For instance, description of the radiomics analysis methodology substantially differs among the scientific articles. In fact, none of the essential methodological elements including imaging protocols, feature extraction, feature selection, and classifier is standardized yet, subsequently resulting in the lack of reproducibility. To take one step forward, extensive investigation into to the technical aspects of the quantification should be accompanied by the clinical significance of the features. Development of standardized CT phantoms that can simulate in vivo measurements or software tools that offer standardized feature processing would be a part of the practical solutions.

Several approaches to neutralize the effect of the reconstruction parameters were described in this review article including resampling, applying conversion factors, or direct image conversion. For the generalization of such approaches, an integrative evaluation of quantitative features should be performed. That is, a single quantitative feature or a single set of measurements for a target disease is not sufficient for the evaluation of adequacy of reconstruction parameter neutralization. Effect of pixel value distribution may vary according to the respective feature classes. Thus, various feature classes have to be analyzed simultaneously.

In addition, a task-based approach should be considered. The effects of reconstruction parameters have little been investigated in task-based manners. Clinical relevance of measurement variability has been reported in terms of Response Evaluation Criteria in Solid

Tumors[74] and low contrast lesion detection.[75, 76] Measurement variability can also be translated into the variation in diagnostic accuracy and reproducibility.[77] Thus, it is recommended that future research include disease-specific tasks (i.e., lung nodule detection or lung cancer diagnosis) to analyze the actual effects of reconstruction parameters or neutralization methodologies.

Conclusion

In conclusion, quantification at chest CT scans is affected by several reconstruction parameters including reconstruction algorithm, kernel, slice thickness and voxel size. Scanner- and software-induced variability are non-negligible, although these are not a part of the reconstruction of the CT images. Several correction or neutralization approaches are being developed and published. Application of deep learning, which can learn differences between inputs and reference data through a residual learning framework, is a promising method as well. Lastly, task-based investigations for the effects of reconstruction parameters and their neutralization methods are recommended for the future research.

References

1. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat Methods Med Res.* 2015;24:9-26.

2. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology standards for quantitative imaging biomarkers. *Radiology.* 2015;277:813-825.

3. Limkin EJ, Sun R, Dercle L, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol.* 2017;28:1191-1206.

4. Kim H, Park CM, Lee M, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One.* 2016;11:e0164924.

5. Solomon J, Mileto A, Nelson RC, et al. Quantitative features of liver lesions, lung nodules, and renal stones at multi-detector row CT examinations: dependency on radiation dose and reconstruction algorithm. *Radiology.* 2016;279:185-194.

6. Lo P, Young S, Kim HJ, et al. Variability in CT lung-nodule quantification: effects of dose reduction and reconstruction methods on density and texture based features. *Med Phys.* 2016;43:4854.

7. Lu L, Ehmke RC, Schwartz LH, et al. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS One.* 2016;11:e0166550.

8. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep.* 2016;6:23428.

9. Zheng Y, Solomon J, Choudhury K, et al. Accuracy and variability of texture-based radiomics features of lung lesions across CT imaging conditions. *SPIE Medical Imaging*. SPIE; 2017;10132:7.

10. He L, Huang Y, Ma Z, et al. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Sci Rep.* 2016;6:34921.

11. Larue R, van Timmeren JE, de Jong EEC, 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;56:1544-1553.

12. Shafiq-Ul-Hassan M, Zhang GG, Hunt DC, et al. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *J Med Imaging (Bellingham).* 2018;5:011013.

13. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS One.* 2017;12:e0178524.

14. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys.* 2017;44:1050-1062.

15. Mackin D, Fave X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol.* 2015;50:757-765.

16. Mahmood U, Apte AP, Deasy JO, et al. Investigating the robustness neighborhood gray tone difference matrix and gray level co-occurrence matrix radiomic features on clinical computed tomography systems using anthropomorphic phantoms: evidence from a multivendor study. *J Comput Assist Tomogr.* 2017;41:995-1001.

17. Yasaka K, Akai H, Mackin D, et al. Precision of quantitative computed tomography texture analysis using image filtering: a phantom study for scanner variability. *Medicine (Baltimore).* 2017;96:e6993.

18. Chen-Mayer HH, Fuld MK, Hoppel B, et al. Standardizing CT lung density measure across scanner manufacturers. *Med Phys.* 2017;44:974-985.

19. Lynch DA, Austin JH, Hogg JC, et al. CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner Society. *Radiology.* 2015;277:192-205.

20. Baumueller S, Hilty R, Nguyen TD, et al. Influence of sinogram-affirmed iterative reconstruction on computed tomography-based lung volumetry and quantification of pulmonary emphysema. *J Comput Assist Tomogr.* 2016;40:96-101.

21. Martin SP, Gariani J, Hachulla AL, et al. Impact of iterative reconstructions on objective and subjective emphysema assessment with computed tomography: a prospective study. *Eur Radiol.* 2017;27:2950-2956.

22. Choo JY, Goo JM, Lee CH, et al. Quantitative analysis of emphysema and airway measurements according to iterative reconstruction algorithms: comparison of filtered back projection, adaptive statistical iterative reconstruction and model-based iterative reconstruction. *Eur Radiol.* 2014;24:799-806.

23. Mets OM, Willemink MJ, de Kort FP, et al. The effect of iterative reconstruction on computed tomography assessment of emphysema, air trapping and airway dimensions. *Eur Radiol.* 2012;22:2103-2109.

24. Nishio M, Koyama H, Ohno Y, et al. Emphysema quantification using ultralow-dose CT with iterative reconstruction and filtered back projection. *AJR Am J Roentgenol.* 2016;206:1184-1192.

25. Yamashiro T, Miyara T, Honda O, et al. Iterative reconstruction for quantitative computed tomography analysis of emphysema: consistent results using different tube currents. *Int J Chron Obstruct Pulmon Dis.* 2015;10:321-327.

26. Rodriguez A, Ranallo FN, Judy PF, et al. The effects of iterative reconstruction and kernel selection on quantitative computed tomography measures of lung density. *Med Phys.* 2017;44:2267-2280.

27. Gierada DS, Bierhals AJ, Choong CK, et al. Effects of CT section thickness and reconstruction kernel on emphysema quantification relationship to the magnitude of the CT emphysema index. *Acad Radiol.* 2010;17:146-156.

28. Hochhegger B, Irion KL, Marchiori E, et al. Reconstruction algorithms influence the follow-up variability in the longitudinal CT emphysema index measurements. *Korean J Radiol.* 2011;12:169-175.

29. Ley-Zaporozhan J, Ley S, Weinheimer O, et al. Quantitative analysis of emphysema in 3D using MDCT: influence of different reconstruction algorithms. *Eur J Radiol.* 2008;65:228-234.

30. Behrendt FF, Das M, Mahnken AH, et al. Computer-aided measurements of pulmonary emphysema in chest multidetector-row spiral computed tomography: effect of image reconstruction parameters. *J Comput Assist Tomogr.* 2008;32:899-904.

31. Boedeker KL, McNitt-Gray MF, Rogers SR, et al. Emphysema: effect of reconstruction algorithm on CT imaging measures. *Radiology.* 2004;232:295-301.

32. Salito C, Woods JC, Aliverti A. Influence of CT reconstruction settings on extremely low attenuation values for specific gas volume calculation in severe emphysema. *Acad Radiol.* 2011;18:1277-1284.

33. Madani A, De Maertelaer V, Zanen J, et al. Pulmonary emphysema: radiation dose and section thickness at multidetector CT quantification--comparison with macroscopic and microscopic morphometry. *Radiology.* 2007;243:250-257.

34. Gallardo-Estrella L, Lynch DA, Prokop M, et al. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. *Eur Radiol.* 2016;26:478-486.

35. Gallardo-Estrella L, Pompe E, de Jong PA, et al. Normalized emphysema scores on

low dose CT: validation as an imaging biomarker for mortality. *PLoS One.* 2017;12:e0188902.

36. Ehrhardt J, Jacob F, Handels H, et al. Comparison of post-hoc normalization approaches for CT-based lung emphysema index quantification. In: Tolxdorff T, Deserno T, Handels H, Meinzer HP, eds. *Bildverarbeitung für die Medizin 2016*. Informatik aktuell. Berlin, Heidelberg: Springer Vieweg; 2016:44-49.

37. Ohkubo M, Wada S, Kayugawa A, et al. Image filtering as an alternative to the application of a different reconstruction kernel in CT imaging: feasibility study in lung cancer screening. *Med Phys.* 2011;38:3915-3923.

38. Jin H, Heo C, Kim JH. Impact of deep learning on the normalization of reconstruction kernel effects in imaging biomarker quantification: a pilot study in CT emphysema. *SPIE Medical Imaging*. SPIE; 2018;10575:7.

39. Wielputz MO, Bardarova D, Weinheimer O, et al. Variation of densitometry on computed tomography in COPD--influence of different software tools. *PLoS One.* 2014;9:e112898.

40. Sciurba FC, Ernst A, Herth FJ, et al. A randomized study of endobronchial valves for advanced emphysema. *N Engl J Med.* 2010;363:1233-1244.

41. Lim HJ, Weinheimer O, Wielputz MO, et al. Fully automated pulmonary lobar segmentation: influence of different prototype software programs onto quantitative evaluation of chronic obstructive lung disease. *PLoS One.* 2016;11:e0151498.

42. Devaraj A, van Ginneken B, Nair A, et al. Use of volumetry for lung nodule management: theory and practice. *Radiology.* 2017;284:630-644.

43. Heuvelmans MA, Walter JE, Vliegenthart R, et al. Disagreement of diameter and volume measurements for pulmonary nodule size estimation in CT lung cancer screening. *Thorax.* 2017 doi: 10.1136/thoraxjnl-2017-210770.

44. Kim H, Park CM. Current perspectives for the size measurement of screening-detected lung nodules. *J Thorac Dis.* 2018;10:1242-1244.

45. den Harder AM, Willemink MJ, van Hamersvelt RW, et al. Pulmonary nodule volumetry at different low computed tomography radiation dose levels with hybrid and model-based iterative reconstruction: a within patient analysis. *J Comput Assist Tomogr.* 2016;40:578-583.

46. Kim H, Park CM, Song YS, et al. Influence of radiation dose and iterative reconstruction algorithms for measurement accuracy and reproducibility of pulmonary nodule volumetry: a phantom study. *Eur J Radiol.* 2014;83:848-857.

47. Siegelman JW, Supanich MP, Gavrielides MA. Pulmonary nodules with ground-glass opacity can be reliably measured with low-dose techniques regardless of iterative reconstruction: results of a phantom study. *AJR Am J Roentgenol.* 2015;204:1242-1247.

48. Wielputz MO, Lederlin M, Wroblewski J, et al. CT volumetry of artificial pulmonary nodules using an ex vivo lung phantom: influence of exposure parameters and iterative reconstruction on reproducibility. *Eur J Radiol.* 2013;82:1577-1583.

49. Willemink MJ, Borstlap J, Takx RA, et al. The effects of computed tomography with iterative reconstruction on solid pulmonary nodule volume quantification. *PLoS One.* 2013;8:e58053.

50. Willemink MJ, Leiner T, Budde RP, et al. Systematic error in lung nodule volumetry: effect of iterative reconstruction versus filtered back projection at different CT parameters. *AJR Am J Roentgenol.* 2012;199:1241-1246.

51. Chen B, Barnhart H, Richard S, et al. Volumetric quantification of lung nodules in CT with iterative reconstruction (ASiR and MBIR). *Med Phys.* 2013;40:111902.

52. Cohen JG, Kim H, Park SB, et al. Comparison of the effects of model-based iterative

reconstruction and filtered back projection algorithms on software measurements in pulmonary subsolid nodules. *Eur Radiol.* 2017;27:3266-3274.

53. Doo KW, Kang EY, Yong HS, et al. Accuracy of lung nodule volumetry in low-dose CT with iterative reconstruction: an anthropomorphic thoracic phantom study. *Br J Radiol.* 2014;87:20130644.

54. Ohno Y, Yaguchi A, Okazaki T, et al. Comparative evaluation of newly developed model-based and commercially available hybrid-type iterative reconstruction methods and filter back projection method in terms of accuracy of computer-aided volumetry (CADv) for low-dose CT protocols in phantom study. *Eur J Radiol.* 2016;85:1375-1382.

55. Sakai N, Yabuuchi H, Kondo M, et al. Volumetric measurement of artificial pure ground-glass nodules at low-dose CT: comparisons between hybrid iterative reconstruction and filtered back projection. *Eur J Radiol.* 2015;84:2654-2662.

56. Ravenel JG, Leue WM, Nietert PJ, et al. Pulmonary nodule volume: effects of reconstruction parameters on automated measurements--a phantom study. *Radiology.* 2008;247:400-408.

57. Ko JP, Rusinek H, Jacobs EL, et al. Small pulmonary nodules: volume measurement at chest CT--phantom study. *Radiology.* 2003;228:864-870.

58. Wang Y, de Bock GH, van Klaveren RJ, et al. Volumetric measurement of pulmonary nodules at low-dose chest CT: effect of reconstruction setting on measurement variability. *Eur Radiol.* 2010;20:1180-1187.

59. Scholten ET, Jacobs C, van Ginneken B, et al. Computer-aided segmentation and volumetry of artificial ground-glass nodules at chest CT. *AJR Am J Roentgenol.* 2013;201:295-300.

60. Christe A, Bronnimann A, Vock P. Volumetric analysis of lung nodules in computed

tomography (CT): comparison of two different segmentation algorithm softwares and two different reconstruction filters on automated volume calculation. *Acta Radiol.* 2014;55:54-61.

61. Zhao B, Schwartz LH, Moskowitz CS, et al. Pulmonary metastases: effect of CT section thickness on measurement--initial experience. *Radiology.* 2005;234:934-939.

62. Petrou M, Quint LE, Nan B, et al. Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and nodule morphology. *AJR Am J Roentgenol.* 2007;188:306-312.

63. Way TW, Chan HP, Goodsitt MM, et al. Effect of CT scanning parameters on volumetric measurements of pulmonary nodules by 3D active contour segmentation: a phantom study. *Phys Med Biol.* 2008;53:1295-1312.

64. Winer-Muram HT, Jennings SG, Meyer CA, et al. Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. *Radiology.* 2003;229:184-194.

65. de Hoop B, Gietema H, van Ginneken B, et al. A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations. *Eur Radiol.* 2009;19:800-808.

66. Ashraf H, de Hoop B, Shaker SB, et al. Lung nodule volumetry: segmentation algorithms within the same software package cannot be used interchangeably. *Eur Radiol.* 2010;20:1878-1885.

67. Zhao YR, van Ooijen PM, Dorrius MD, et al. Comparison of three software systems for semi-automatic volumetry of pulmonary nodules on baseline and follow-up CT examinations. *Acta Radiol.* 2014;55:691-698.

68. Kim H, Park CM, Lee SM, et al. A comparison of two commercial volumetry software programs in the analysis of pulmonary ground-glass nodules: segmentation capability and

measurement accuracy. *Korean J Radiol.* 2013;14:683-691.

69. Kalpathy-Cramer J, Zhao B, Goldgof D, et al. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. *J Digit Imaging.* 2016;29:476-487.

70. Rinaldi MF, Bartalena T, Braccaioli L, et al. Three-dimensional analysis of pulmonary nodules: variability of semiautomated volume measurements between different versions of the same software. *Radiol Med.* 2010;115:403-412.

71. Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax.* 2015;70 Suppl 2:ii1-ii54.

72. Horeweg N, van der Aalst CM, Vliegenthart R, et al. Volumetric computed tomography screening for lung cancer: three rounds of the NELSON trial. *Eur Respir J.* 2013;42:1659-1667.

73. Xu DM, Gietema H, de Koning H, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer.* 2006;54:177-184.

74. Oxnard GR, Zhao B, Sima CS, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *J Clin Oncol.* 2011;29:3114-3119.

75. Christianson O, Chen JJ, Yang Z, et al. An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations. *Radiology.* 2015;275:725-734.

76. Saiprasad G, Filliben J, Peskin A, et al. Evaluation of low-contrast detectability of iterative reconstruction across multiple institutions, CT scanner manufacturers, and radiation exposure levels. *Radiology.* 2015;277:124-133.

77. Kim H, Park CM, Hwang EJ, et al. Pulmonary subsolid nodules: value of semi-automatic measurement in diagnostic accuracy, diagnostic reproducibility and nodule

classification agreement. *Eur Radiol.* 2018;28:2124-2133.

78. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology*. 2018;288:407-15.

Figure Legends

FIGURE 1. An example of a radiomics analysis work flow.

Commented [C6]: Reviewer 2-1

After image acquisition, feature extraction including intensity, shape, and texture features is performed. Then, feature selection and redundancy reduction procedure are executed. Various classifiers (regression, clustering, random forest, or support vector machines) can be used for the radiomics model building. This image was reproduced from a study by Berenguer et al.[78] Copyright belongs to the Radiological Society of North America and the authors of the original article.

FIGURE ~~1~~2. Influence of iterative reconstruction on the radiomics features.

Features were obtained from 69 patients with pulmonary subsolid nodules who underwent CT scans of a uniform protocol. Kernel density plots were drawn for (A) GLCM entropy and (B) GLCM contrast to investigate their distribution. It can be recognized that the distribution of radiomics features differs according to the reconstruction algorithms and this may have clinical impact for the diagnostic or prognostic modeling based on these features.

FBP, filtered back projection; GLCM, gray-level co-occurrence matrix; MBIR, model-based iterative reconstruction

FIGURE 3. Effect of reconstruction kernel on the radiomics features

Commented [C7]: Reviewer 2-2

Radiomics features (n=125) including shape, intensity, and texture features (GLCM and NGTDM) were extracted from the two sets of chest CT scans reconstructed with (A) chest kernel and (B) standard kernel for squamous cell carcinoma in right lower lobe. (C) A waterfall plot demonstrating coefficients variation of radiomics features between the reconstruction

kernels in descending order (the feature with the highest variability at the left side). The coefficient of variation ranged from 0% to 33%. GLCM contrast showed the highest variability and intensity features were robust to the kernel. Shape features were equivalent between the two kernels as the same region-of-interest was applied for the feature extraction.

GLCM, gray-level co-occurrence matrix; NGTDM, neighborhood gray-tone difference matrix

FIGURE 4. Noise reducing strength of IR algorithms and emphysema index calculation.

Commented [C8]: Reviewer 2-2

(A) Iterative Model Reconstruction (Philips Healthcare), a knowledge-based IR algorithm which accounts knowledge of data statistics, image statistics, and system models, showed higher level of noise reduction than (B) iDose, a hybrid IR. Subsequently, the measured emphysema index was lower at IMR (C; 13.5%) than at iDose (D; 16.9%).

FIGURE ~~2~~5. Effect of reconstruction kernels on the emphysema quantification.

A non-enhanced CT scan from a patient with chronic obstructive pulmonary disease was reconstructed with (A) soft kernel (B; Philips Healthcare) and (B) sharp kernel (YD). Fully automated quantification of percentage low attenuation area (≤950 Hounsfield Unit) was conducted (C, soft kernel; D, sharp kernel) and it was substantially different between the two kernels, which was 18.0% at the soft kernel and 32.7% at the sharp kernel. Thus, reconstruction kernel is not interchangeable for the emphysema quantification.

FIGURE ~~3~~6. Impact of reconstruction algorithms on the lung nodule segmentation.

A part-solid nodule was segmented using a commercial semi-automatic volumetry software (AVIEW, Coreline Soft, Seoul, Korea) for a CT scan reconstructed with (A) FBP and (B) MBIR. Volume of the total nodule and its solid portion were (C) 758.1 mm^3 and 98.0 mm^3, respectively, at FBP and (D) 729.2 mm^3 and 103.0 mm^3, respectively, at MBIR. Thus, volumetric

measurements were comparable between the reconstruction algorithms. Note that segmentation parameters (shape and attenuation threshold) were adjusted respectively at both reconstruction algorithms and that the outer border was underestimated for both FBP and MBIR.

FBP, filtered back projection; MBIR, model-based iterative reconstruction

Effect of reconstruction parameters on the quantitative analysis of chest computed tomography

Summary

Quantitative features obtained from CT scans are being explored for clinical applications. Various classes of quantitative features exist for chest CT including radiomics features, emphysema measurements, lung nodule volumetric measurements, dual energy quantification, and perfusion parameters. A number of research articles have shown promise in diagnosis and prognosis prediction of oncologic patients or those with diffuse lung diseases using these feature classes. Nevertheless, a prerequisite for the quantification is the evaluation of variation in measurements in terms of repeatability and reproducibility, which are distinct aspects of precision but are often not separable from each other. There are well-known sources of measurement variability including patient factors, CT acquisition (scan and reconstruction) factors, and radiologist (or measurement-related) factors. The purpose of this article is to review the effects of CT reconstruction parameters on the quantitative imaging features and efforts to correct or neutralize variations induced by those parameters.

Key Words

computed tomography; reconstruction; radiomics; emphysema; volumetry; quantification

Introduction

Quantitative computed tomography (CT) features vary from simple uni-dimensional measurements to those calculated from complex three-dimensional (3D) matrices of pixel value distribution. Although deep learning is beyond the scope of this article, it is also a way of automatically extracting and learning a number of quantitative readouts from simple to complex features. Ideally, a quantitative imaging readout can be used as an imaging biomarker, which is defined as an objectively measured characteristic derived from an in vivo image as an indicator of normal biologic processes, pathogenic processes, or response to a therapeutic intervention.[1]

Before the clinical application of any quantitative features, linearity, bias, and precision have to be scrutinized.[2] Evaluation of these statistical and metrologic methodologies is essential for the translation of research-level quantification to clinical practice. According to Sullivan et al.,[2] linearity is the ability to provide measured values that are directly proportional to the true values. Bias, commonly termed as accuracy, is the difference between the mean of measurements determined from the same object and true value.[2] Precision is variability of the measurement, which can be classified into repeatability and reproducibility.[2] Repeatability refers to the variation of feature values at the repeated identical imaging condition, while reproducibility is the variation of the measurement according to the different conditions (i.e., CT scanners, protocols, and institutions).[2]

Reproducibility of measurements has been studied extensively in the field of radiology as the range of variation is crucial to determine the true biological change in vivo and to determine the normal range of observations. There are a lot of potential sources of variations including imaging acquisition factors (technician, scanner, manufacturer, acquisition

parameters, and reconstruction parameters) and radiologist factors (operator and software tool). Any of these factors can affect the reproducibility of measurements.

The importance of reproducibility especially at a longitudinal study is that measurement change can be attributable to either true change or measurement variation. If the measurement variation is large, the clinical decision cannot be made with confidence. In addition, if inter-individual dynamic range (measurement differences between individuals) of a certain feature is smaller than the potential measurement variation, then that feature may not be utilized as an imaging biomarker. Nevertheless, conversely, it means that recognition of imaging factors that generate variation raises an issue to standardize quantification.

In this review, we aimed to describe implications of several CT reconstruction parameters on the quantitative imaging features from the aspect of measurement reproducibility. We dealt with radiomics features, emphysema quantification and lung nodule volumetry. Influence of reconstruction parameters such as iterative reconstruction algorithms, reconstruction kernels, and slice thickness will be discussed.

Radiomics features

Radiomics analysis refers to high-throughput extraction of quantitative features including intensity or texture features from images. It is usually performed in a following order: lesion segmentation, feature extraction, feature selection, model training and validation (Fig. 1). Radiomics analysis enables to capture and calculate independent imaging features (e.g. tumor heterogeneity) which may or may not be visible to the human eyes. Radiomics features include first-order features which do not consider spatial relationships among voxel values and second-order features from gray-level co-occurrence matrix (GLCM), gray-level run length matrix, gray-level size zone matrix, and neighborhood gray-tone difference matrix (NGTDM). This method has potential to further promote the role of imaging in the era of precision medicine.[3] A number of studies based on the radiomics analysis for diagnosis, cancer staging, prognosis prediction, treatment response monitoring, and surveillance have been reported to date.[3] Nevertheless, standardization of analysis protocols, reproducibility of features, and redundancy of extracted information are the remaining concerns for the radiomics approach.

Effect of iterative reconstruction algorithm

A few studies have shown the effect of iterative reconstruction (IR) on the radiomics features (Fig. 2).[4-6] Kim et al.[4] reported that most of the first-order intensity features and second-order GLCM-based features showed significant differences between filtered back projection (FBP) and Sinogram Affirmed Iterative Reconstruction (SAFIRE) using chest CT scans of patients with pulmonary tumors. Interestingly, features were also substantially influenced by the noise reduction strength of SAFIRE (level 3 vs. 5). Size features, entropy, and GLCM entropy were the most robust features (coefficient of variation ≤5%) when inter-reader variability in tumor segmentation and inter-reconstruction algorithm variability were

taken into account. Solomon et al.[5] also revealed that five first-order features and two GLCM-based features were affected by model-based iterative reconstruction (MBIR). These studies showed that the feature values differed significantly according to the reconstruction algorithm of CT scans and the range of variation could be substantial. However, none of those analyzed the potentially substantive clinical implication of the measurement variability induced by IR in terms of diagnosis or prognostication. Measurement variation of features may affect the actual performance of a diagnostic or prognostic model by erroneously increasing the overlap between the classes of label data. This topic warrants future investigations.

Effect of reconstruction kernel and slice thickness

As for the effect of reconstruction kernel (Fig. 3) and slice thickness on the radiomics features, Lu et al.[7] analyzed concordance correlation coefficients (CCCs) of the radiomics features between reconstruction settings of different slice thickness (1.25, 2.5, and 5 mm) and kernels (lung and standard kernel). They found that the agreement levels of changing reconstruction kernels were lower than those of changing slice thickness. Obviously, changing both the reconstruction kernel and the slice thickness (slice thickness of 1.25 mm and lung kernel vs. slice thickness of 5mm and standard kernel) resulted in the worst agreement (CCCs <0.51). Among multiple radiomics features, size, mean density, coarse boundary morphology, and coarse texture features were relatively robust to the changes in slice thickness and kernels. Intriguingly, boundary sharpness and fine texture features, which contained detailed morphological or textural information, were vulnerable to the reconstruction settings. Zhao et al.[8] performed a similar study using same-day repeat CT scans. They revealed that the radiomics features obtained from thin-section (1.25 mm) images had higher inter-scan agreement. Laplacian of Gaussian features obtained from the standard kernel were more

reproducible than those from the lung kernel. Zheng et al.[9] performed a phantom study with 3D printed textured lesions. They concluded that among multiple variables of radiation dosage (0.67, 1.42, and 5.80 mGy), reconstruction algorithm (FBP and IR), kernel (standard, soft, and edge), and slice thickness (0.6 and 5 mm), the last two parameters were most influential. In addition, features from thin-slice and edge reconstruction kernel were more accurate and reproducible. The discrepancy of the study results between Zhao et al.[8] and Zheng et al.[9] (standard vs. edge kernel for the reproducibility) might be due to differences in extracted radiomics features. Some features might be more vulnerable to the sharp kernel, while others might show higher variation to the standard kernel.

A task-based assessment for the implication of the reconstruction parameters was performed by He et al.[10] Their investigation demonstrated a link between a variation in feature values and the actual diagnosis. They conducted regression analysis with the radiomics features obtained from several different CT scans (enhanced vs. non-enhanced; 1.25 vs. 5 mm; standard vs. lung kernel) for the differentiation of malignant lung nodules. They concluded that the optimal combination for the highest diagnostic performance was non-contrast, thin-slice, and standard kernel. The combination of enhancement status and reconstruction parameters had substantial effects on the feature selection and subsequently on the performance of diagnostic models.

There have been attempts to reduce the variation of radiomics features. Larue et al.[11] focused on the resampling method to reduce the variability or dependency of radiomics features on the slice thickness. They found that most radiomics features were affected by the slice thickness and this variation could be reduced by resampling of voxel sizes using cubic or linear interpolation before feature extraction. They also demonstrated that linear interpolation (voxel

size resampling into 1x1x3 mm^3) resulted in feature stability in 48% of the radiomics features. For the kernel-induced variability, a recent study[12] suggested that 3D noise power spectrum peak frequency and region of interest maximum intensity could be used as correction factors.

Effect of pixel size

Given the fact that slice thickness causes variation in the radiomics features, it is easily understandable that the effect of pixel size or in-plane resolution on the radiomics features would also be considerable. Mackin et al.[13] compared the intra-patient variability caused by the variation in pixel size using CT scans reconstructed with 5 different pixel sizes ranging from 0.59 to 0.98 mm. They retrospectively reconstructed CT scans of 8 lung cancer patients (tumor size not reported) with varying field-of-views from 30 to 50 cm in 5 cm increments. The intra-patient variability (overall CCC) was larger than the inter-patient variability in 79% of the features, which implied that the variation caused by the pixel size may obscure true inter-individual differences. For the correction of variation in pixel sizes, they combined image resampling (bilinear interpolation to 1 mm/pixel) with Butterworth low-pass filtering in the frequency domain. After correction, the intra-patient variability was relatively large in only 10% of the features. Shafiq-ul-Hassan et al.[14] reported a similar finding. Among 213 features, variation in 42 features diminished significantly after resampling (percent coefficient of variation <30). In their study, 150 features were reproducible irrespective of the voxel sizes and 21 had substantial variation before and after voxel size resampling.

Effect of CT scanner

CT scanner manufacturer and model is another cause of variability for the radiomics features. Mackin et al.[15] performed 17 CT scans using CT scanners from four manufacturers. Although the scanning parameters were not equal across multiple protocols and CT scanners,

the interesting finding was that the radiomics feature values obtained from the same vendor were grouped together at hierarchical clustering. For instance, entropy or strength values obtained from Philips scanners showed negative deviation from the normalized mean of 0 and those from Siemens and Toshiba (currently, Canon Medical Systems) scanners exhibited positive deviation from the normalized mean. Variation due to CT scanners was observed even after acquiring CT scans with near-identical scanning parameters.[16] Mahmood et al.[16] obtained NGTDM and GLCM features from three CT scanners. Scanning parameters including voxel size, radiation dosage, pitch, and kernels were matched across the CT scans. However, NGTDM and GLCM features were not reproducible among different scanners with CCCs of less than 0.9. Scanner-induced variability was also reported for the image filtration-based features.[17] Scanner-induced variation is probably due to detector design, beam spectra, calibration methods, and quality control/maintenance of CT machines.

A robust correction factor or development of a less variable feature, which is clinically relevant at the same time, would be required. From this aspect, Chen-Mayer et al.[18] proposed a 5-step calibration procedure using a single parameter to describe scanner dependent contribution to the pixel value. They mapped CT numbers to 80 keV and demonstrated that reproducibility of the pixel values markedly improved after calibration (standard deviation less than 1 Hounsfield Unit [HU]). As radiomic features were not evaluated in this study, further investigation using the described calibration procedure is required.

Emphysema quantification

CT provides valuable quantitative parameters for the prediction of the degree of airflow obstruction at pulmonary function testing and the risk of chronic obstructive pulmonary disease (COPD) exacerbations.[19] CT quantification for COPD evaluation can be categorized into emphysema quantification such as percentage low-attenuation area (%LAA; sometimes termed as relative area -950 HU or emphysema index), airway measurement (airway wall thickness, wall area, lumen diameter, or internal perimeter) and air trapping analysis. These quantitative features have been investigated in terms of reproducibility.

Effect of iterative reconstruction algorithm

Studies to date have consistently shown that IR has significant influence on the quantification of LAA. A number of commercial IR algorithms have been assessed including SAFIRE from Siemens Healthcare;[20, 21] adaptive statistical iterative reconstruction (ASIR) and MBIR from GE Healthcare;[21, 22] iDose from Philips Healthcare;[23] and adaptive iterative dose reduction using 3D processing (AIDR3D) from Canon Medical Systems.[24, 25] These referenced studies demonstrated that LAA would be underestimated if IR is applied. IR changes the distribution of pixel values in the extremes of the attenuation histogram.[20] That is, the mean of pixel value is preserved, but the standard deviation is reduced and the density histogram becomes more sharply peaked.[26] Such convergence toward the mean leads to decrease in the number of pixels below a certain cutoff such as -950 HU. In addition, such phenomenon is accentuated when the IR strength or noise reduction level increase.[20, 21] This means that the grade of denoising affects LAA measurement. Underestimation of LAA was greater with MBIR than with ASIR (Fig. 4).[21, 22]

A beneficial effect of IR in emphysema quantification was reported by Japanese researchers.[24, 25] They suggested that the agreement of LAA between standard-dose and low-dose (or ultra-low-dose) CT scans could be improved by using IR.[24, 25] That is, increased image noise in low-dose CT, which may result in an increase in LAA (overestimation), could be mitigated by the application of IR. This finding is promising given that the effect of noise on quantification, such as in cases of low-dose setting or scanning large patients, can be relieved by IR. Another potential strength of IR was proposed by Choo et al.[22] They performed a phantom study using a CTP674 Lung Phantom which included polycarbonate tubes simulating human airways. They compared the accuracy of airway measurements between the reconstruction algorithms (FBP, ASIR, and MBIR). Notably, MBIR exhibited the most accurate results for the measurements of luminal area, wall area, and wall thickness.

A more challenging method for the preservation of pixel value distribution of FBP after application of IR was suggested by Rodriguez et al.[26] They compared pixel value distribution of phantom inserts among CT scan combinations of different kernels (standard and sharp) and reconstruction algorithms (FBP, ASIR, and MBIR). In their study, combination of bone kernel with ASIR had similar pixel value distribution with the pair of standard kernel and FBP. It should be noted that various emphysema quantification parameters should be tested to guarantee the comparability of feature values between that combination and this approach should be analyzed for other CT scanners and vendors.

Effect of reconstruction kernel and slice thickness

It is now a well-known fact that the reconstruction kernel should be kept constant for the longitudinal emphysema quantification. A number of studies have demonstrated that application of a sharp kernel can cause an erroneous increase of LAA (Fig. 5).[27-30] Sharp kernel

modifies HU of interface pixels between structures with significantly different attenuation coefficients.[31] Therefore, detection of tissue or material interface can be affected by the kernel used.[28] In addition, increased image noise associated with sharp kernel broadens the width of pixel value histogram, which subsequently results in increased LAA as abovementioned. The effect of slice thickness can be explained in the same way. Thin-slice reconstruction can also cause increment in LAA.[27, 30, 32, 33]

Madani et al.[33] reported correlations between LAA from smooth kernel-reconstructed CT scan and macroscopic and microscopic morphometric measurements. Hochhegger et al.[28] proposed that a standard kernel should be preferred given the larger longitudinal variability of LAA at sharp kernel CT scans. Gierada et al.[27] also presented the pros of smooth kernel based on the stronger correlation of emphysema index with histological alveolar wall distance measurements at smooth kernels.

To relieve the variability in emphysema quantification caused by the reconstruction kernel, Gallardo-Estrella et al.[34] suggested a normalization method of different reconstruction kernels by frequency band decomposition with hierarchical unsharp masking to standardize the energy in each band to a reference value. By using this method, calculated energy coefficients can be applied to various kernel images to create normalized images. This method effectively reduced variation in emphysema quantification caused by the kernel. Gallardo-Estrella et al.[35] also reported that emphysema quantification after pixel resampling to 3 mm slice thickness, normalization, and bullae analysis to minimize variability in slice thickness, kernel, and noise led to better prognostication of all-cause mortality and lung cancer mortality. Another group of researchers compared three kernel normalization methods and reported that edge-preserving frequency decomposition was the best normalization method for quantification of emphysema

index.[36] Ohkubo et al.[37] reported that the ratio of modulation transfer functions for the two different kernels could be used as a filter function for kernel conversion. However, emphysema quantification was not investigated in this study. More recently, deep learning was applied for the kernel conversion from sharp to smooth for the quantification of LAA.[38] Pairs of standard and sharp kernel images were fed to convolutional network for training and the deep learning model produced LAA of converted images (8.87 ± 6.20%, converted B50f; 27.65 ± 7.28%, original B50f) similar to that of the standard kernel images (10.82 ± 6.71%, B30f).[38]

Effect of quantification software

Another potential source of variation in emphysema measurement is the software program for automated quantification, which causes variation in lung segmentation, airway segmentation and subsequent quantification.[39] Wielputz et al.[39] compared three software tools including one in-house software and two commercial software products from major vendors. They revealed that the inter-software variability of emphysema index was higher than the measurable progression with median differences from -5.0 to -1.7%. Limits of agreement were as wide as -25.5 to 18.8%.[39] They also stated that such a large variability range could affect patient inclusion or exclusion for endobronchial valve treatment.[40] Therefore, the researchers described that the software cannot be used interchangeably for the longitudinal follow-up or post-treatment evaluation.[39] Inter-software variability for the lobe-based quantification was reported by Lim et al.[41] Four fully automated lobar segmentation tools were analyzed and limits of agreement for emphysema index was as large as -7 to 14%.[41] Patients with inhomogeneous emphysema distribution, who are suitable for surgical or bronchoscopic lung volume reduction surgery, showed higher inter-software variability due to greater distortion of normal anatomy.[41]

Lung nodule volumetry

Volumetric measurement has a relatively long history compared to other quantitative features, which goes back to the late 1990s. During the past decades, its technical as well as clinical performance have been investigated by researchers.[42] Volume measurement has several merits over diameter measurement. First, volume is more representative of the true dimension of a nodule considering substantial diameter variation within a nodule.[43] Second, it has potential to further stratify the intermediate risk category which is determined by diameter measurement.[43] Third, semi-automated volumetry may provide reproducible measurements.[44]

Effect of iterative reconstruction algorithm

To date, most studies have reported that the effect of IR on the lung nodule volumetry was not clinically relevant.[45-50] In detail, there was either no statistically significant difference in volumetric results between FBP and IR or the magnitude of IR-induced variation was smaller than the inter-reader or inter-scan measurement variability (Fig. 6). Extensive studies have been performed, which investigated the effect of IR with and without other factors such as radiation dose on the lung nodule volumetry in terms of bias and precision. IR algorithms including ASIR and MBIR;[51, 52] SAFIRE;[48] AIDR 3D and forward projected model-based iterative reconstruction (FIRST; Canon Medical Systems]);[53, 54] iDose and IMR[45-47, 49, 50, 55] were investigated using phantoms or in vivo lung nodules. Considering that the inherent contrast between the pulmonary nodules and background parenchyma is high, it can be expected that the margin segmentation might not be much affected by the reconstruction algorithms. For the phantom nodules, researchers have shown that even the volumetric measurements of subsolid nodules, were robust to the change in the reconstruction algorithms.[46, 47] Regarding the subsolid nodules in vivo, Cohen et al.[52] compared the volumetric parameters between FBP and MBIR

and found that volume and mass of the subsolid nodules as a whole and their solid components were measured significantly larger when using MBIR.[52] Nevertheless, differences between the reconstruction algorithms (up to 6.6%) were within the range of intra- and inter-reader variability.[52] A few studies suggested that the measurement accuracy for the volumetric analysis of subsolid nodules was higher with IR at extremely low radiation dose levels (10-20 mA).[53-55]

Effect of reconstruction kernel

The effect of the reconstruction kernel on lung nodule volumetry is somewhat controversial. Ravanel et al.[56] reported that lung kernel yielded the least bias among seven different reconstruction kernels but they also stated that no single kernel consistently yielded both the least biased and the most precise volume estimations. Ko et al.[57] also reported that a high-frequency kernel provided less bias than a low-frequency algorithm. However, Wang et al.[58] performed three repeated measurements of solid lung nodules to analyze repeatability and found that soft kernel was better than sharp kernel, particularly for non-smooth-round nodules. For the artificial subsolid nodules, Scholten et al.[59] reported that absolute percentage error was not influenced by the kernel. More recently, Christe et al.[60] compared soft and sharp kernels for 113 lung nodules and reported that the mean volume was measured generally higher using the soft kernel. The relative difference of volume between the kernels ranged from 5.9% to 20.5% depending on the software programs.[60] Limits of agreement between the kernels for one of the tools ranged from -7.9% to 49%.[60]

Based on these studies, it is difficult to choose an optimal kernel setting for the lung nodule volumetry as both bias and precision are the essential components of measurements. However, it is obvious that consistency in reconstruction kernel is definitely required given the

substantial difference of volume measurement according to the kernel selection. Further research for the recent scanners and software programs based on both phantom and in vivo nodules are warranted for conclusive results.

Effect of slice thickness

Slice thickness particularly matters for the small lung nodules as the proportion of surface voxels increases in the small nodules.[42] If section thickness increases, the margin of nodules becomes blurred due to partial volume averaging of surface voxels.[42] Accordingly, the estimated volume would increase. Volume overestimation at thick-section CT scans was reported by Ravanel et al.[56] using synthetic nodules and Zhao et al.[61] using in vivo metastatic nodules. Other researchers reported similar findings.[62, 63] In addition, Wang et al.[58] demonstrated that slice thickness was associated with repeatability of measurements and the relative volume differences in 2 mm thickness (22.5%) were greater than those in 1mm thickness (8.9%). Voxel resampling as mentioned in the previous section would be helpful for the longitudinal follow-up of nodules on CT scans with different slice thicknesses. The error correction equation can also be applied.[64]

Effect of volumetry software programs

General consensus from the published articles is that the software program or its internal segmentation algorithms have a substantial effect on the measured nodule volume.[65-70] It is to be noted that the software-induced variation can sometimes be greater than the nodule volume growth cutoff (25%),[66, 67] which is currently used at the British Thoracic Society guideline[71] or Dutch-Belgian lung cancer screening trial (NELSON).[72] This indicates that the same software or segmentation algorithm should be consistently used for the surveillance of nodules in order not to misinterpret measurement variation as true growth. This would be

particularly important for the lung cancer screening setting or treatment response evaluation. Zhao et al.[67] performed volume doubling time (VDT) calculation using CT scans of three time points with three software tools. They reported that the consensus on VDT categorization at follow-up was less than 50% when the results from different software programs were compared.[67] In addition, Kalpathy-Cramer et al.[69] revealed that the most accurate segmentation algorithm was not the most repeatable. Therefore, care should be taken for the selection of a segmentation algorithm based on the thorough evaluation of both bias and precision. This is important as the screening-detected nodules are categorized by the baseline nodule volume and also classified by VDT at follow-up scans.[73]

Further directions of reproducibility analysis

Many studies for the quantitative imaging biomarkers have been conducted thus far. However, most of them are still research-based and most are not ready for actual clinical application. One of the main reasons is the standardization of methodology. For instance, description of the radiomics analysis methodology substantially differs among the scientific articles. In fact, none of the essential methodological elements including imaging protocols, feature extraction, feature selection, and classifier is standardized yet, subsequently resulting in the lack of reproducibility. To take one step forward, extensive investigation into to the technical aspects of the quantification should be accompanied by the clinical significance of the features. Development of standardized CT phantoms that can simulate in vivo measurements or software tools that offer standardized feature processing would be a part of the practical solutions.

Several approaches to neutralize the effect of the reconstruction parameters were described in this review article including resampling, applying conversion factors, or direct image conversion. For the generalization of such approaches, an integrative evaluation of quantitative features should be performed. That is, a single quantitative feature or a single set of measurements for a target disease is not sufficient for the evaluation of adequacy of reconstruction parameter neutralization. Effect of pixel value distribution may vary according to the respective feature classes. Thus, various feature classes have to be analyzed simultaneously.

In addition, a task-based approach should be considered. The effects of reconstruction parameters have little been investigated in task-based manners. Clinical relevance of measurement variability has been reported in terms of Response Evaluation Criteria in Solid

Tumors[74] and low contrast lesion detection.[75, 76] Measurement variability can also be translated into the variation in diagnostic accuracy and reproducibility.[77] Thus, it is recommended that future research include disease-specific tasks (i.e., lung nodule detection or lung cancer diagnosis) to analyze the actual effects of reconstruction parameters or neutralization methodologies.

Conclusion

In conclusion, quantification at chest CT scans is affected by several reconstruction parameters including reconstruction algorithm, kernel, slice thickness and voxel size. Scanner- and software-induced variability are non-negligible, although these are not a part of the reconstruction of the CT images. Several correction or neutralization approaches are being developed and published. Application of deep learning, which can learn differences between inputs and reference data through a residual learning framework, is a promising method as well. Lastly, task-based investigations for the effects of reconstruction parameters and their neutralization methods are recommended for the future research.

References

1. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat Methods Med Res.* 2015;24:9-26.

2. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology standards for quantitative imaging biomarkers. *Radiology.* 2015;277:813-825.

3. Limkin EJ, Sun R, Dercle L, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol.* 2017;28:1191-1206.

4. Kim H, Park CM, Lee M, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One.* 2016;11:e0164924.

5. Solomon J, Mileto A, Nelson RC, et al. Quantitative features of liver lesions, lung nodules, and renal stones at multi-detector row CT examinations: dependency on radiation dose and reconstruction algorithm. *Radiology.* 2016;279:185-194.

6. Lo P, Young S, Kim HJ, et al. Variability in CT lung-nodule quantification: effects of dose reduction and reconstruction methods on density and texture based features. *Med Phys.* 2016;43:4854.

7. Lu L, Ehmke RC, Schwartz LH, et al. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS One.* 2016;11:e0166550.

8. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep.* 2016;6:23428.

9. Zheng Y, Solomon J, Choudhury K, et al. Accuracy and variability of texture-based radiomics features of lung lesions across CT imaging conditions. *SPIE Medical Imaging*. SPIE; 2017;10132:7.

10. He L, Huang Y, Ma Z, et al. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Sci Rep.* 2016;6:34921.

11. Larue R, van Timmeren JE, de Jong EEC, 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;56:1544-1553.

12. Shafiq-Ul-Hassan M, Zhang GG, Hunt DC, et al. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *J Med Imaging (Bellingham).* 2018;5:011013.

13. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS One.* 2017;12:e0178524.

14. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys.* 2017;44:1050-1062.

15. Mackin D, Fave X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol.* 2015;50:757-765.

16. Mahmood U, Apte AP, Deasy JO, et al. Investigating the robustness neighborhood gray tone difference matrix and gray level co-occurrence matrix radiomic features on clinical computed tomography systems using anthropomorphic phantoms: evidence from a multivendor study. *J Comput Assist Tomogr.* 2017;41:995-1001.

17. Yasaka K, Akai H, Mackin D, et al. Precision of quantitative computed tomography texture analysis using image filtering: a phantom study for scanner variability. *Medicine (Baltimore).* 2017;96:e6993.

18. Chen-Mayer HH, Fuld MK, Hoppel B, et al. Standardizing CT lung density measure across scanner manufacturers. *Med Phys.* 2017;44:974-985.

19. Lynch DA, Austin JH, Hogg JC, et al. CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner Society. *Radiology.* 2015;277:192-205.

20. Baumueller S, Hilty R, Nguyen TD, et al. Influence of sinogram-affirmed iterative reconstruction on computed tomography-based lung volumetry and quantification of pulmonary emphysema. *J Comput Assist Tomogr.* 2016;40:96-101.

21. Martin SP, Gariani J, Hachulla AL, et al. Impact of iterative reconstructions on objective and subjective emphysema assessment with computed tomography: a prospective study. *Eur Radiol.* 2017;27:2950-2956.

22. Choo JY, Goo JM, Lee CH, et al. Quantitative analysis of emphysema and airway measurements according to iterative reconstruction algorithms: comparison of filtered back projection, adaptive statistical iterative reconstruction and model-based iterative reconstruction. *Eur Radiol.* 2014;24:799-806.

23. Mets OM, Willemink MJ, de Kort FP, et al. The effect of iterative reconstruction on computed tomography assessment of emphysema, air trapping and airway dimensions. *Eur Radiol.* 2012;22:2103-2109.

24. Nishio M, Koyama H, Ohno Y, et al. Emphysema quantification using ultralow-dose CT with iterative reconstruction and filtered back projection. *AJR Am J Roentgenol.* 2016;206:1184-1192.

25. Yamashiro T, Miyara T, Honda O, et al. Iterative reconstruction for quantitative computed tomography analysis of emphysema: consistent results using different tube currents. *Int J Chron Obstruct Pulmon Dis.* 2015;10:321-327.

26. Rodriguez A, Ranallo FN, Judy PF, et al. The effects of iterative reconstruction and kernel selection on quantitative computed tomography measures of lung density. *Med Phys.* 2017;44:2267-2280.

27. Gierada DS, Bierhals AJ, Choong CK, et al. Effects of CT section thickness and reconstruction kernel on emphysema quantification relationship to the magnitude of the CT emphysema index. *Acad Radiol.* 2010;17:146-156.

28. Hochhegger B, Irion KL, Marchiori E, et al. Reconstruction algorithms influence the follow-up variability in the longitudinal CT emphysema index measurements. *Korean J Radiol.* 2011;12:169-175.

29. Ley-Zaporozhan J, Ley S, Weinheimer O, et al. Quantitative analysis of emphysema in 3D using MDCT: influence of different reconstruction algorithms. *Eur J Radiol.* 2008;65:228-234.

30. Behrendt FF, Das M, Mahnken AH, et al. Computer-aided measurements of pulmonary emphysema in chest multidetector-row spiral computed tomography: effect of image reconstruction parameters. *J Comput Assist Tomogr.* 2008;32:899-904.

31. Boedeker KL, McNitt-Gray MF, Rogers SR, et al. Emphysema: effect of reconstruction algorithm on CT imaging measures. *Radiology.* 2004;232:295-301.

32. Salito C, Woods JC, Aliverti A. Influence of CT reconstruction settings on extremely low attenuation values for specific gas volume calculation in severe emphysema. *Acad Radiol.* 2011;18:1277-1284.

33. Madani A, De Maertelaer V, Zanen J, et al. Pulmonary emphysema: radiation dose and section thickness at multidetector CT quantification--comparison with macroscopic and microscopic morphometry. *Radiology.* 2007;243:250-257.

34. Gallardo-Estrella L, Lynch DA, Prokop M, et al. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. *Eur Radiol.* 2016;26:478-486.

35. Gallardo-Estrella L, Pompe E, de Jong PA, et al. Normalized emphysema scores on

low dose CT: validation as an imaging biomarker for mortality. *PLoS One.* 2017;12:e0188902.

36. Ehrhardt J, Jacob F, Handels H, et al. Comparison of post-hoc normalization approaches for CT-based lung emphysema index quantification. In: Tolxdorff T, Deserno T, Handels H, Meinzer HP, eds. *Bildverarbeitung für die Medizin 2016*. Informatik aktuell. Berlin, Heidelberg: Springer Vieweg; 2016:44-49.

37. Ohkubo M, Wada S, Kayugawa A, et al. Image filtering as an alternative to the application of a different reconstruction kernel in CT imaging: feasibility study in lung cancer screening. *Med Phys.* 2011;38:3915-3923.

38. Jin H, Heo C, Kim JH. Impact of deep learning on the normalization of reconstruction kernel effects in imaging biomarker quantification: a pilot study in CT emphysema. *SPIE Medical Imaging*. SPIE; 2018;10575:7.

39. Wielputz MO, Bardarova D, Weinheimer O, et al. Variation of densitometry on computed tomography in COPD--influence of different software tools. *PLoS One.* 2014;9:e112898.

40. Sciurba FC, Ernst A, Herth FJ, et al. A randomized study of endobronchial valves for advanced emphysema. *N Engl J Med.* 2010;363:1233-1244.

41. Lim HJ, Weinheimer O, Wielputz MO, et al. Fully automated pulmonary lobar segmentation: influence of different prototype software programs onto quantitative evaluation of chronic obstructive lung disease. *PLoS One.* 2016;11:e0151498.

42. Devaraj A, van Ginneken B, Nair A, et al. Use of volumetry for lung nodule management: theory and practice. *Radiology.* 2017;284:630-644.

43. Heuvelmans MA, Walter JE, Vliegenthart R, et al. Disagreement of diameter and volume measurements for pulmonary nodule size estimation in CT lung cancer screening. *Thorax.* 2017 doi: 10.1136/thoraxjnl-2017-210770.

44. Kim H, Park CM. Current perspectives for the size measurement of screening-detected lung nodules. *J Thorac Dis.* 2018;10:1242-1244.

45. den Harder AM, Willemink MJ, van Hamersvelt RW, et al. Pulmonary nodule volumetry at different low computed tomography radiation dose levels with hybrid and model-based iterative reconstruction: a within patient analysis. *J Comput Assist Tomogr.* 2016;40:578-583.

46. Kim H, Park CM, Song YS, et al. Influence of radiation dose and iterative reconstruction algorithms for measurement accuracy and reproducibility of pulmonary nodule volumetry: a phantom study. *Eur J Radiol.* 2014;83:848-857.

47. Siegelman JW, Supanich MP, Gavrielides MA. Pulmonary nodules with ground-glass opacity can be reliably measured with low-dose techniques regardless of iterative reconstruction: results of a phantom study. *AJR Am J Roentgenol.* 2015;204:1242-1247.

48. Wielputz MO, Lederlin M, Wroblewski J, et al. CT volumetry of artificial pulmonary nodules using an ex vivo lung phantom: influence of exposure parameters and iterative reconstruction on reproducibility. *Eur J Radiol.* 2013;82:1577-1583.

49. Willemink MJ, Borstlap J, Takx RA, et al. The effects of computed tomography with iterative reconstruction on solid pulmonary nodule volume quantification. *PLoS One.* 2013;8:e58053.

50. Willemink MJ, Leiner T, Budde RP, et al. Systematic error in lung nodule volumetry: effect of iterative reconstruction versus filtered back projection at different CT parameters. *AJR Am J Roentgenol.* 2012;199:1241-1246.

51. Chen B, Barnhart H, Richard S, et al. Volumetric quantification of lung nodules in CT with iterative reconstruction (ASiR and MBIR). *Med Phys.* 2013;40:111902.

52. Cohen JG, Kim H, Park SB, et al. Comparison of the effects of model-based iterative

reconstruction and filtered back projection algorithms on software measurements in pulmonary subsolid nodules. *Eur Radiol.* 2017;27:3266-3274.

53. Doo KW, Kang EY, Yong HS, et al. Accuracy of lung nodule volumetry in low-dose CT with iterative reconstruction: an anthropomorphic thoracic phantom study. *Br J Radiol.* 2014;87:20130644.

54. Ohno Y, Yaguchi A, Okazaki T, et al. Comparative evaluation of newly developed model-based and commercially available hybrid-type iterative reconstruction methods and filter back projection method in terms of accuracy of computer-aided volumetry (CADv) for low-dose CT protocols in phantom study. *Eur J Radiol.* 2016;85:1375-1382.

55. Sakai N, Yabuuchi H, Kondo M, et al. Volumetric measurement of artificial pure ground-glass nodules at low-dose CT: comparisons between hybrid iterative reconstruction and filtered back projection. *Eur J Radiol.* 2015;84:2654-2662.

56. Ravenel JG, Leue WM, Nietert PJ, et al. Pulmonary nodule volume: effects of reconstruction parameters on automated measurements--a phantom study. *Radiology.* 2008;247:400-408.

57. Ko JP, Rusinek H, Jacobs EL, et al. Small pulmonary nodules: volume measurement at chest CT--phantom study. *Radiology.* 2003;228:864-870.

58. Wang Y, de Bock GH, van Klaveren RJ, et al. Volumetric measurement of pulmonary nodules at low-dose chest CT: effect of reconstruction setting on measurement variability. *Eur Radiol.* 2010;20:1180-1187.

59. Scholten ET, Jacobs C, van Ginneken B, et al. Computer-aided segmentation and volumetry of artificial ground-glass nodules at chest CT. *AJR Am J Roentgenol.* 2013;201:295-300.

60. Christe A, Bronnimann A, Vock P. Volumetric analysis of lung nodules in computed

tomography (CT): comparison of two different segmentation algorithm softwares and two different reconstruction filters on automated volume calculation. *Acta Radiol.* 2014;55:54-61.

61. Zhao B, Schwartz LH, Moskowitz CS, et al. Pulmonary metastases: effect of CT section thickness on measurement--initial experience. *Radiology.* 2005;234:934-939.

62. Petrou M, Quint LE, Nan B, et al. Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and nodule morphology. *AJR Am J Roentgenol.* 2007;188:306-312.

63. Way TW, Chan HP, Goodsitt MM, et al. Effect of CT scanning parameters on volumetric measurements of pulmonary nodules by 3D active contour segmentation: a phantom study. *Phys Med Biol.* 2008;53:1295-1312.

64. Winer-Muram HT, Jennings SG, Meyer CA, et al. Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. *Radiology.* 2003;229:184-194.

65. de Hoop B, Gietema H, van Ginneken B, et al. A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations. *Eur Radiol.* 2009;19:800-808.

66. Ashraf H, de Hoop B, Shaker SB, et al. Lung nodule volumetry: segmentation algorithms within the same software package cannot be used interchangeably. *Eur Radiol.* 2010;20:1878-1885.

67. Zhao YR, van Ooijen PM, Dorrius MD, et al. Comparison of three software systems for semi-automatic volumetry of pulmonary nodules on baseline and follow-up CT examinations. *Acta Radiol.* 2014;55:691-698.

68. Kim H, Park CM, Lee SM, et al. A comparison of two commercial volumetry software programs in the analysis of pulmonary ground-glass nodules: segmentation capability and

measurement accuracy. *Korean J Radiol.* 2013;14:683-691.

69. Kalpathy-Cramer J, Zhao B, Goldgof D, et al. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. *J Digit Imaging.* 2016;29:476-487.

70. Rinaldi MF, Bartalena T, Braccaioli L, et al. Three-dimensional analysis of pulmonary nodules: variability of semiautomated volume measurements between different versions of the same software. *Radiol Med.* 2010;115:403-412.

71. Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax.* 2015;70 Suppl 2:ii1-ii54.

72. Horeweg N, van der Aalst CM, Vliegenthart R, et al. Volumetric computed tomography screening for lung cancer: three rounds of the NELSON trial. *Eur Respir J.* 2013;42:1659-1667.

73. Xu DM, Gietema H, de Koning H, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer.* 2006;54:177-184.

74. Oxnard GR, Zhao B, Sima CS, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *J Clin Oncol.* 2011;29:3114-3119.

75. Christianson O, Chen JJ, Yang Z, et al. An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations. *Radiology.* 2015;275:725-734.

76. Saiprasad G, Filliben J, Peskin A, et al. Evaluation of low-contrast detectability of iterative reconstruction across multiple institutions, CT scanner manufacturers, and radiation exposure levels. *Radiology.* 2015;277:124-133.

77. Kim H, Park CM, Hwang EJ, et al. Pulmonary subsolid nodules: value of semi-automatic measurement in diagnostic accuracy, diagnostic reproducibility and nodule

classification agreement. *Eur Radiol.* 2018;28:2124-2133.

78. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology*. 2018;288:407-15.

Figure Legends

FIGURE 1. An example of a radiomics analysis work flow.

After image acquisition, feature extraction including intensity, shape, and texture features is performed. Then, feature selection and redundancy reduction procedure are executed. Various classifiers (regression, clustering, random forest, or support vector machines) can be used for the radiomics model building. This image was reproduced from a study by Berenguer et al.[78] Copyright belongs to the Radiological Society of North America and the authors of the original article.

FIGURE 2. Influence of iterative reconstruction on the radiomics features.

Features were obtained from 69 patients with pulmonary subsolid nodules who underwent CT scans of a uniform protocol. Kernel density plots were drawn for (A) GLCM entropy and (B) GLCM contrast to investigate their distribution. It can be recognized that the distribution of radiomics features differs according to the reconstruction algorithms and this may have clinical impact for the diagnostic or prognostic modeling based on these features.

FBP, filtered back projection; GLCM, gray-level co-occurrence matrix; MBIR, model-based iterative reconstruction

FIGURE 3. Effect of reconstruction kernel on the radiomics features

Radiomics features (n=125) including shape, intensity, and texture features (GLCM and NGTDM) were extracted from the two sets of chest CT scans reconstructed with (A) chest kernel and (B) standard kernel for squamous cell carcinoma in right lower lobe. (C) A waterfall plot demonstrating coefficients variation of radiomics features between the reconstruction

kernels in descending order (the feature with the highest variability at the left side). The coefficient of variation ranged from 0% to 33%. GLCM contrast showed the highest variability and intensity features were robust to the kernel. Shape features were equivalent between the two kernels as the same region-of-interest was applied for the feature extraction.

GLCM, gray-level co-occurrence matrix; NGTDM, neighborhood gray-tone difference matrix

FIGURE 4. Noise reducing strength of IR algorithms and emphysema index calculation.

(A) Iterative Model Reconstruction (Philips Healthcare), a knowledge-based IR algorithm which accounts knowledge of data statistics, image statistics, and system models, showed higher level of noise reduction than (B) iDose, a hybrid IR. Subsequently, the measured emphysema index was lower at IMR (C; 13.5%) than at iDose (D; 16.9%).

FIGURE 5. Effect of reconstruction kernels on the emphysema quantification.

A non-enhanced CT scan from a patient with chronic obstructive pulmonary disease was reconstructed with (A) soft kernel (B; Philips Healthcare) and (B) sharp kernel (YD). Fully automated quantification of percentage low attenuation area (≤950 Hounsfield Unit) was conducted (C, soft kernel; D, sharp kernel) and it was substantially different between the two kernels, which was 18.0% at the soft kernel and 32.7% at the sharp kernel. Thus, reconstruction kernel is not interchangeable for the emphysema quantification.

FIGURE 6. Impact of reconstruction algorithms on the lung nodule segmentation.

A part-solid nodule was segmented using a commercial semi-automatic volumetry software (AVIEW, Coreline Soft, Seoul, Korea) for a CT scan reconstructed with (A) FBP and (B) MBIR. Volume of the total nodule and its solid portion were (C) 758.1 mm^3 and 98.0 mm^3, respectively, at FBP and (D) 729.2 mm^3 and 103.0 mm^3, respectively, at MBIR. Thus, volumetric

measurements were comparable between the reconstruction algorithms. Note that segmentation parameters (shape and attenuation threshold) were adjusted respectively at both reconstruction algorithms and that the outer border was underestimated for both FBP and MBIR.

FBP, filtered back projection; MBIR, model-based iterative reconstruction

Figure 1

Click here to access/download;Figure;Fig 1.tif

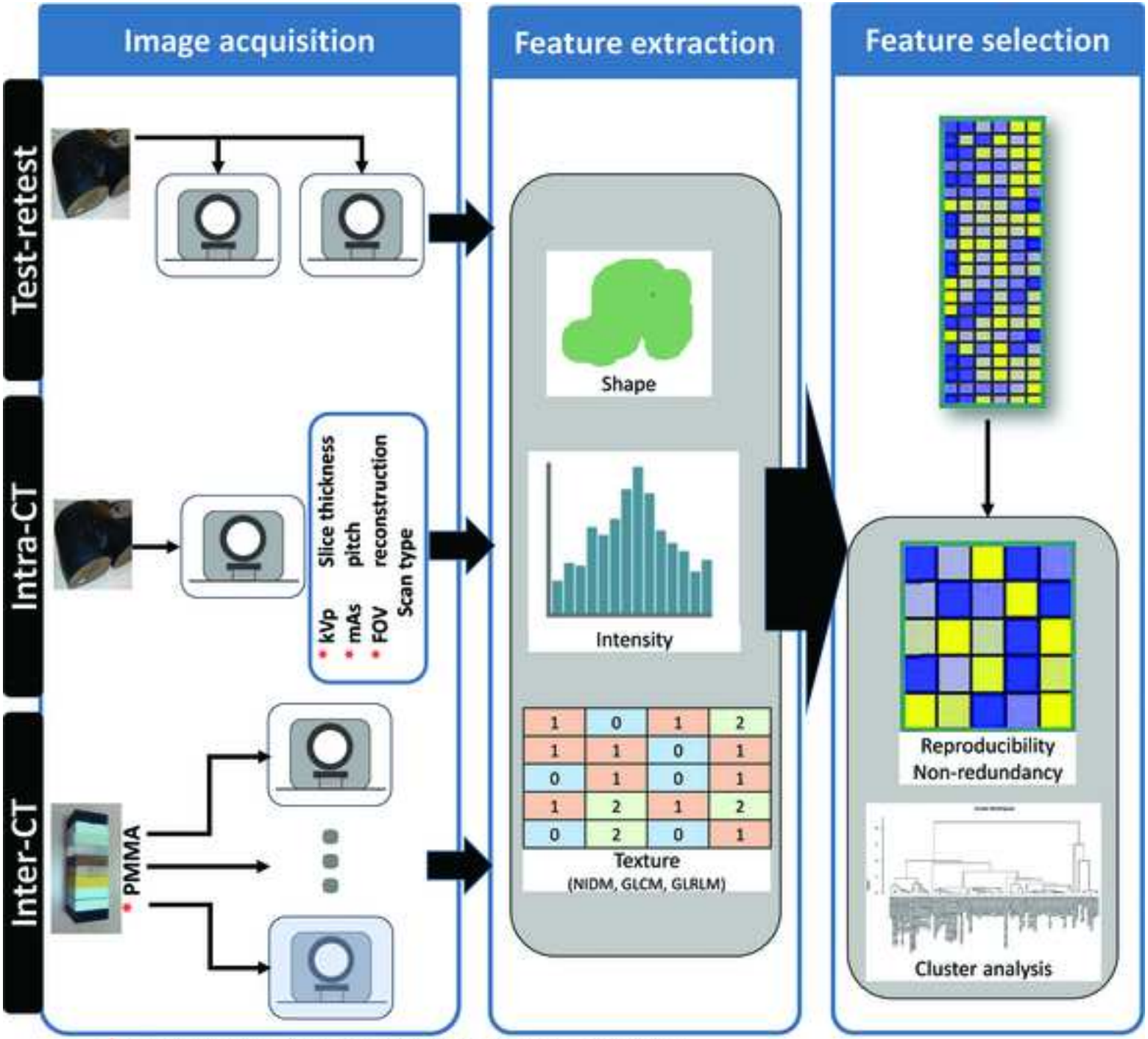

Figure 2A

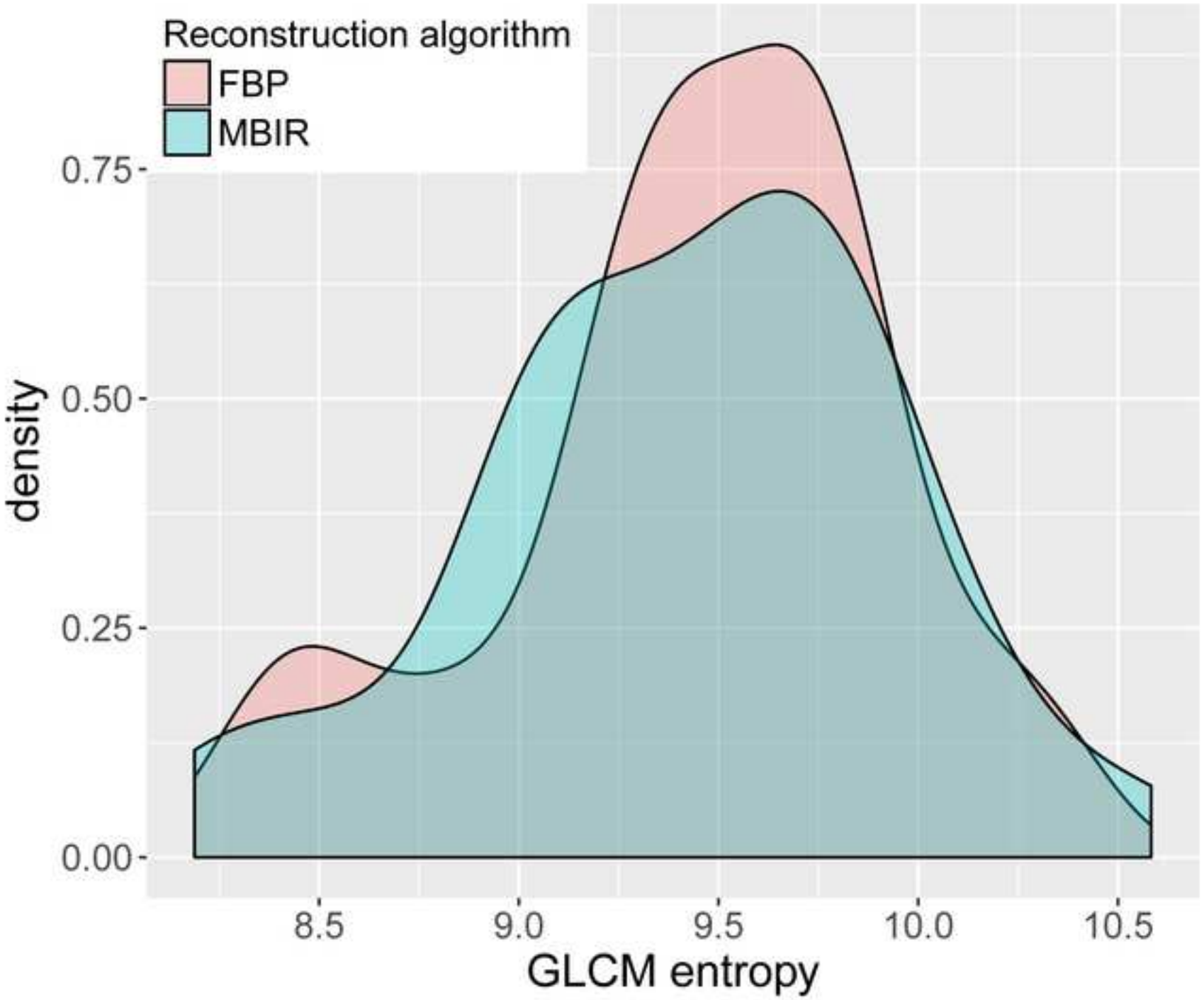

Figure 2B

Click here to access/download;Figure;Fig 2B.tif

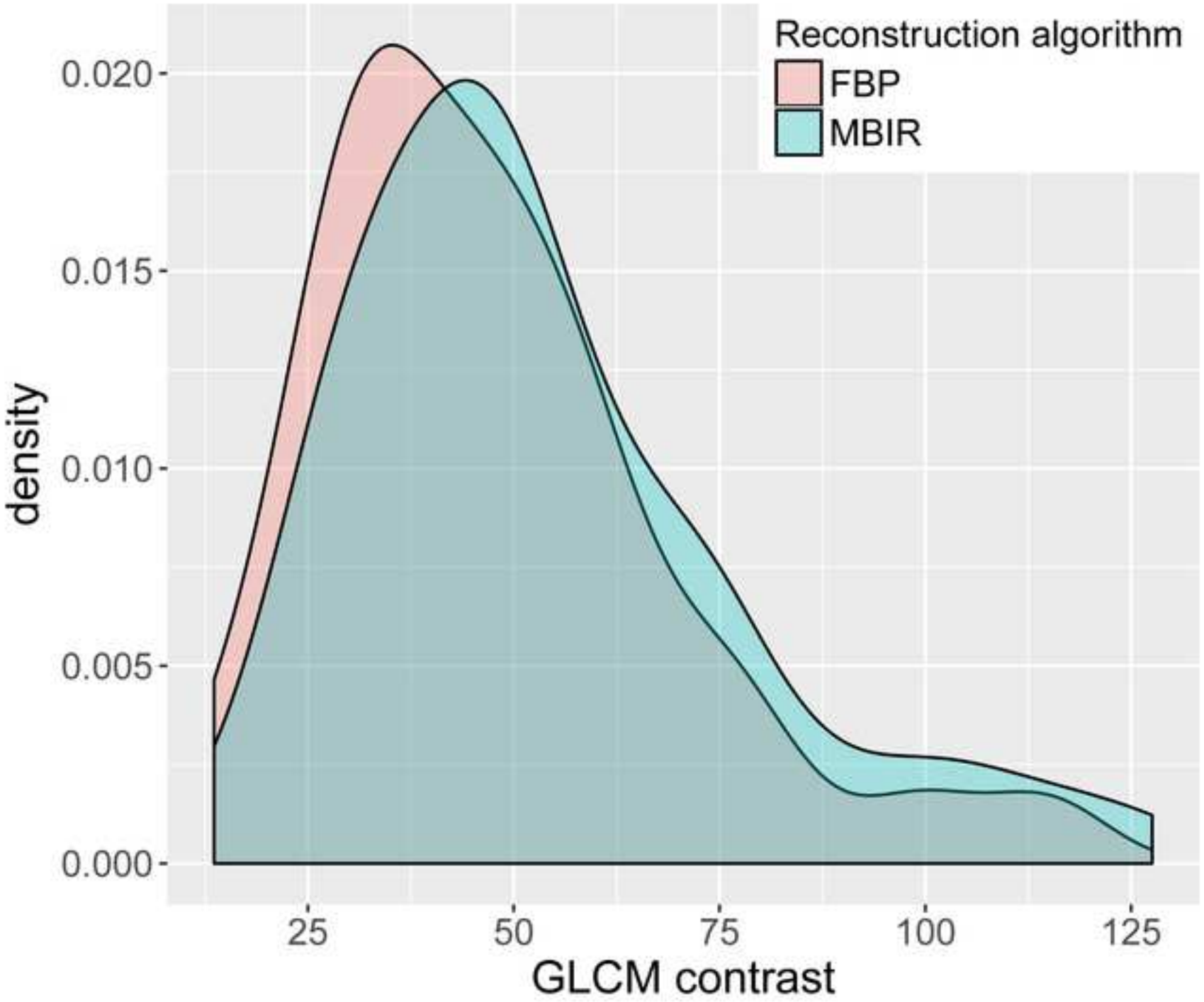

Figure 3A

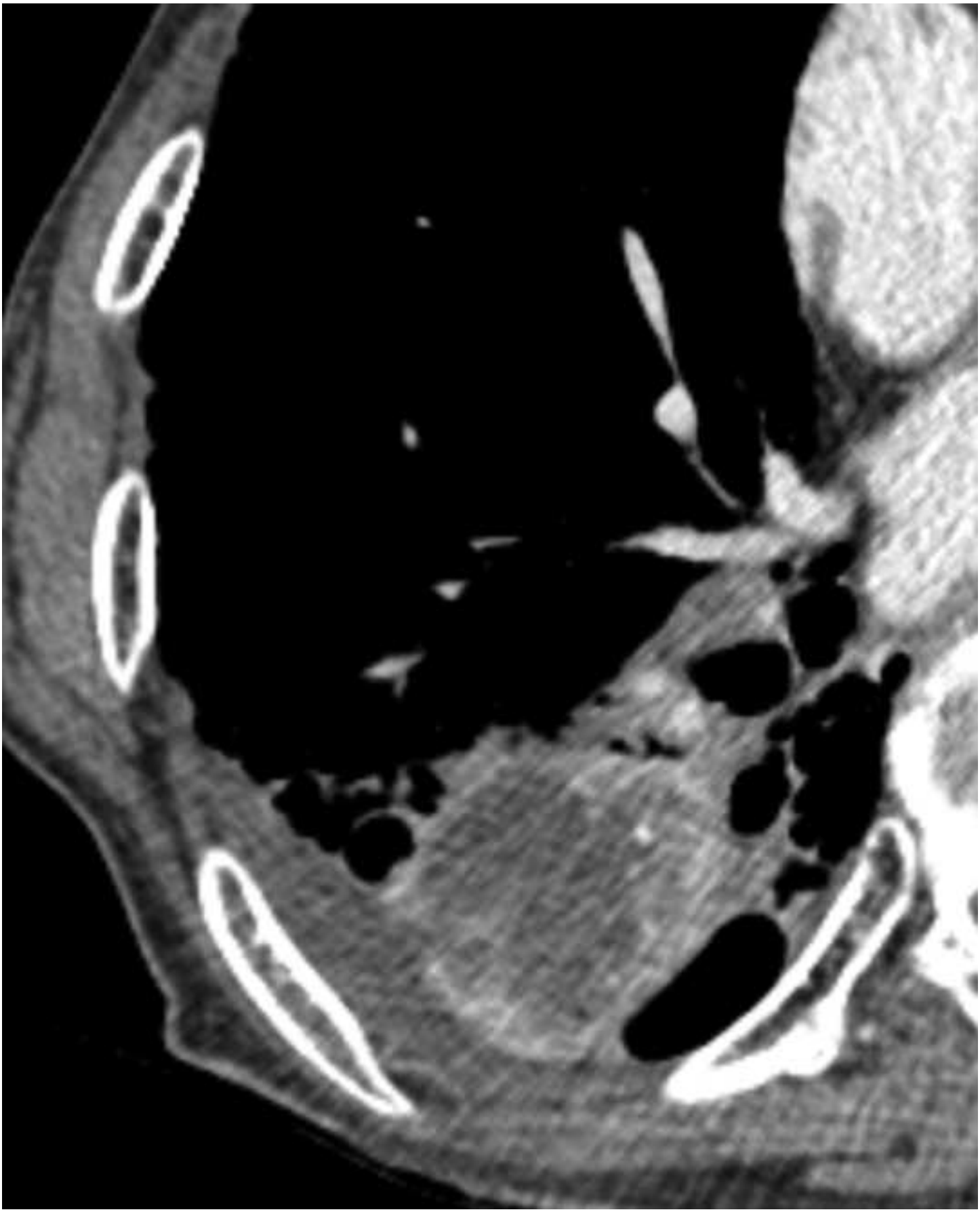

Figure 3B

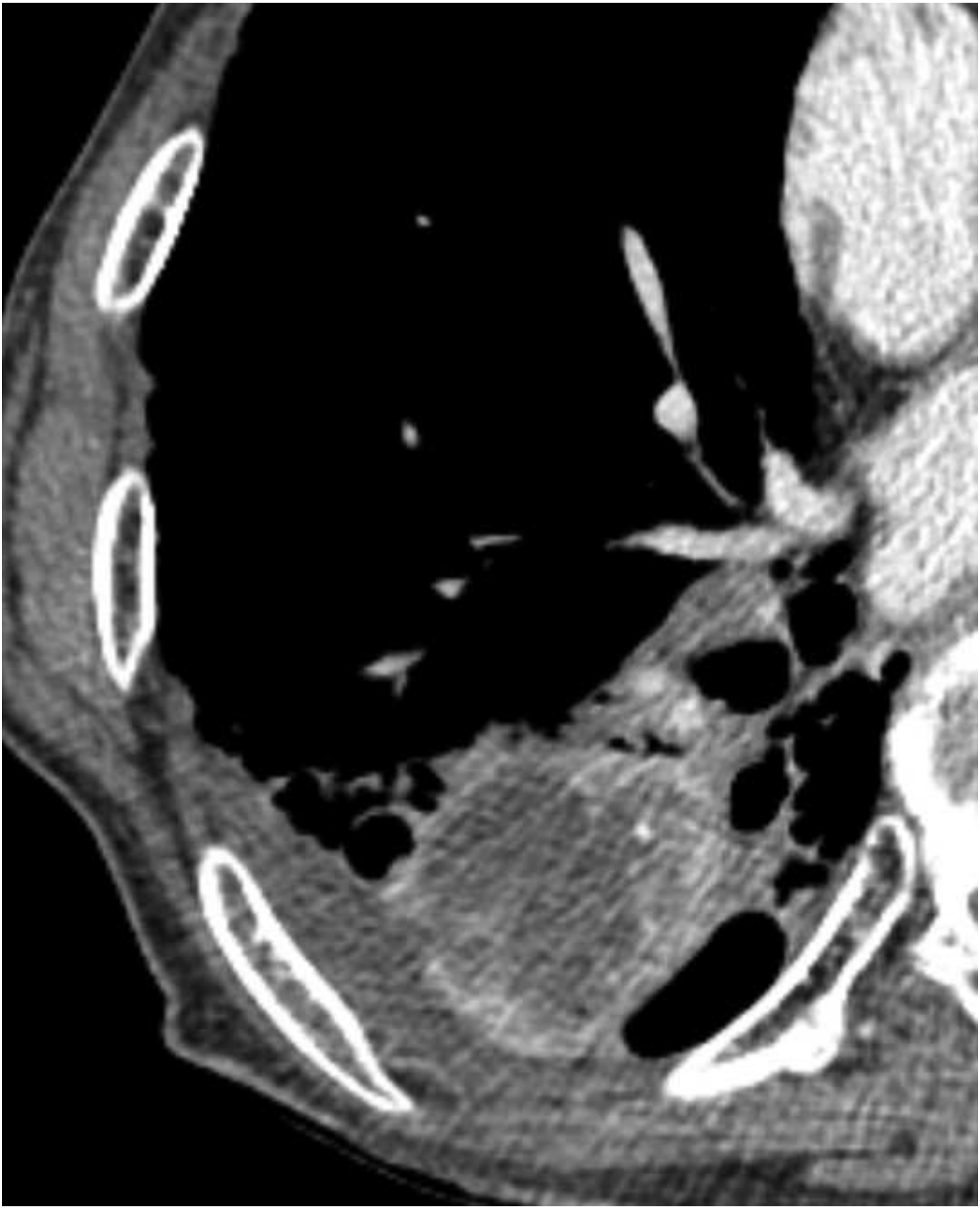

Figure 3C

Click here to access/download;Figure;Fig 3C.tif

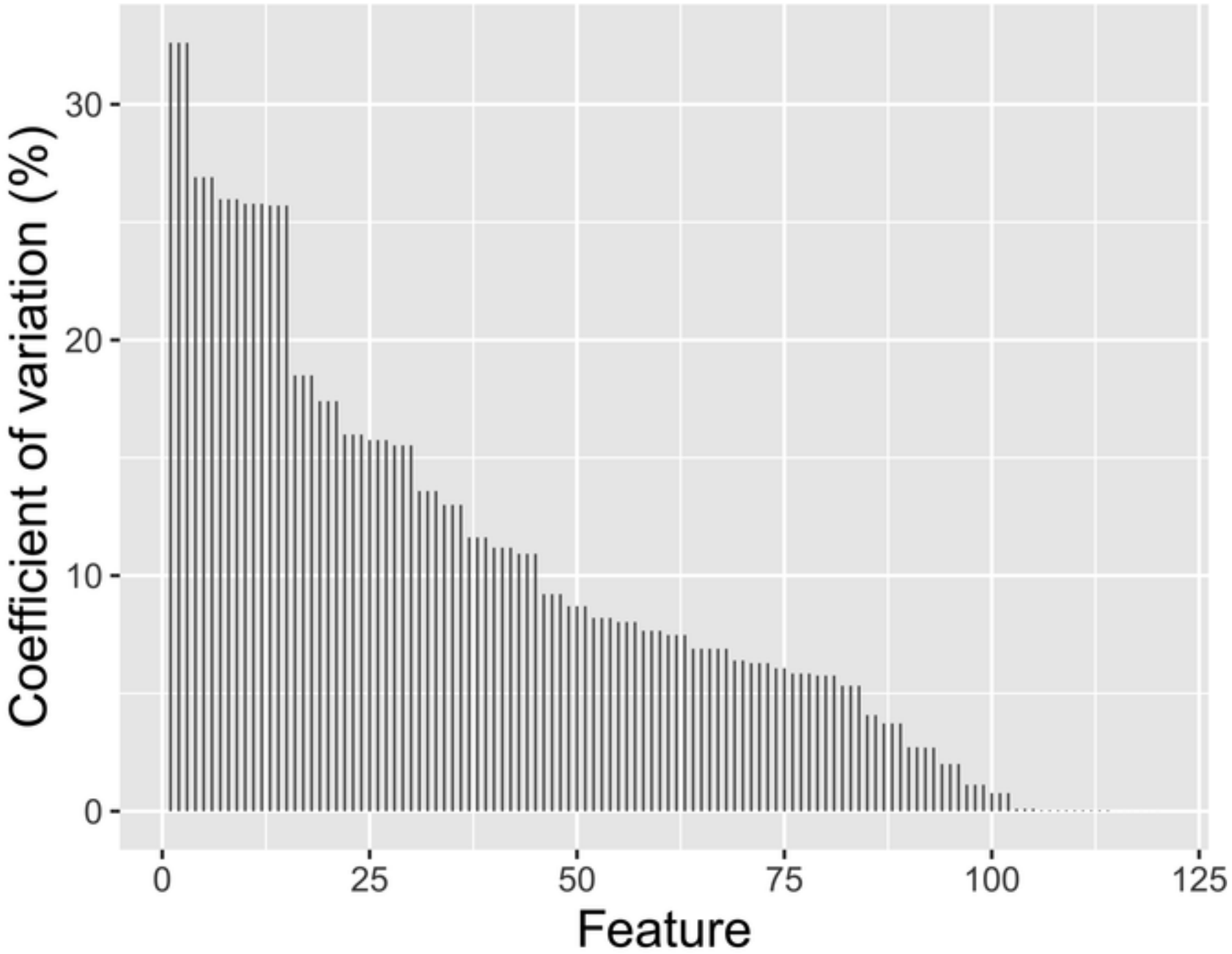

Figure 4A

Click here to access/download;Figure;Fig 4A.tif

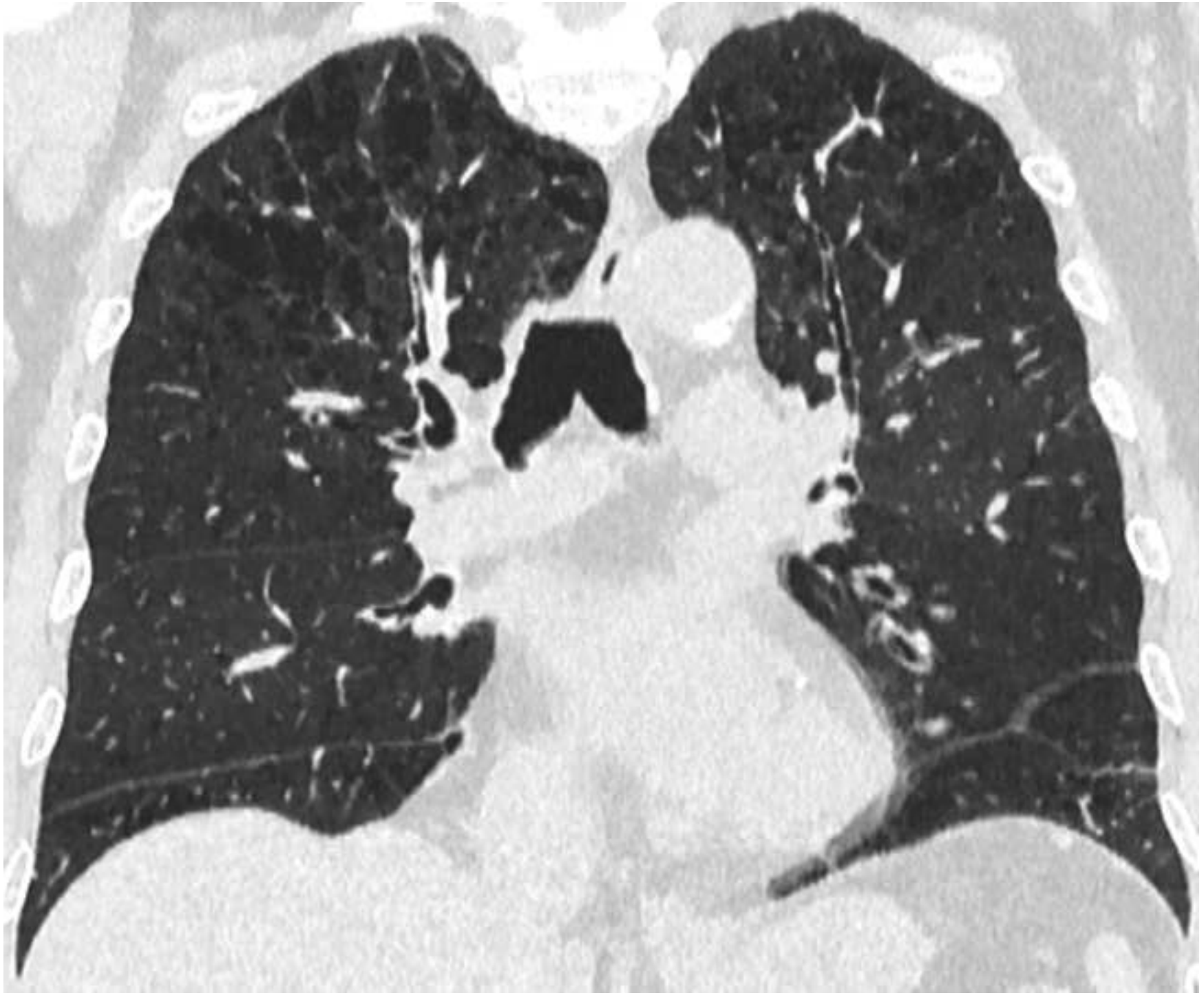

Figure 4B

Click here to access/download;Figure;Fig 4B.tif

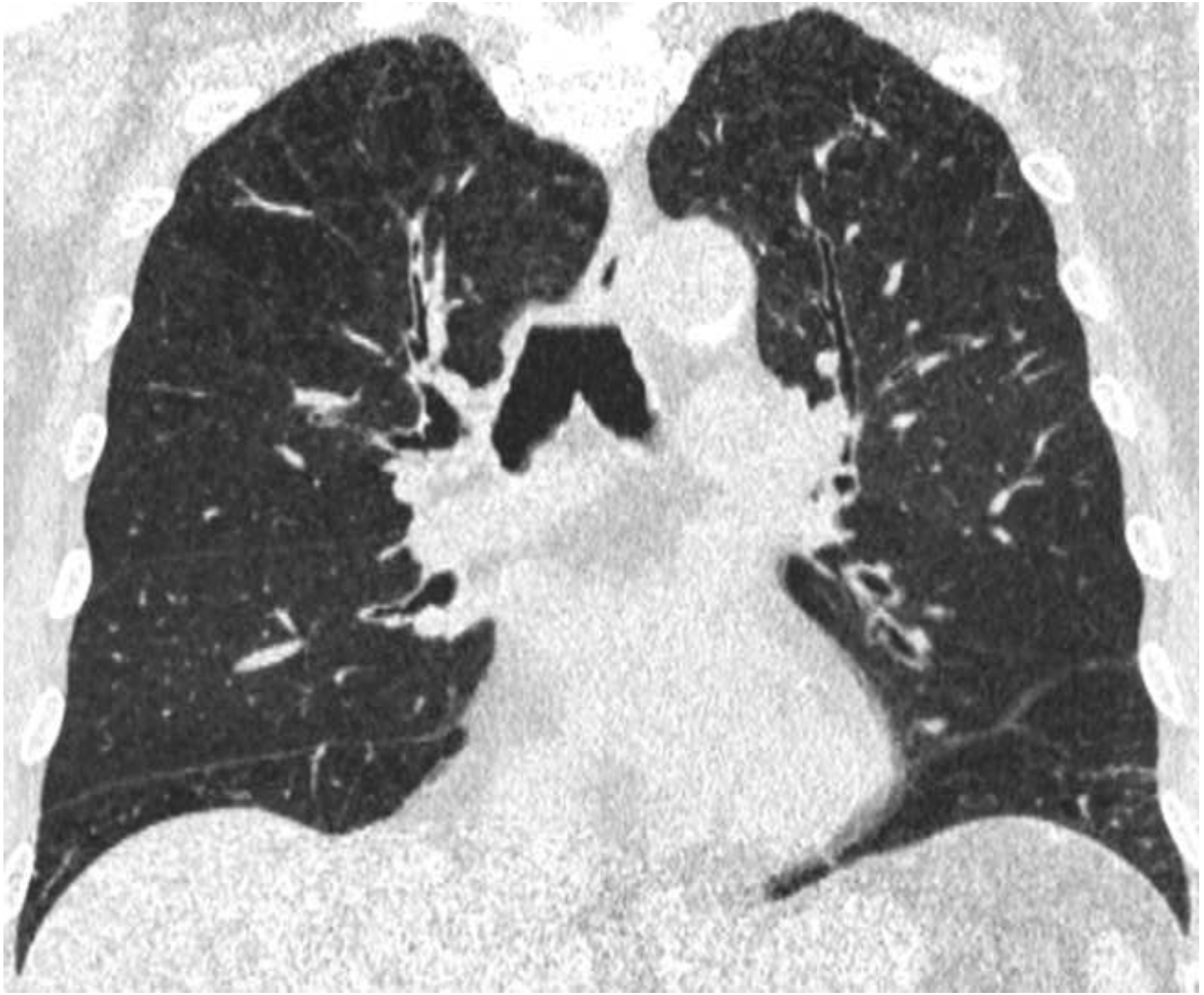

Figure 4C

Click here to access/download;Figure;Fig 4C.tif

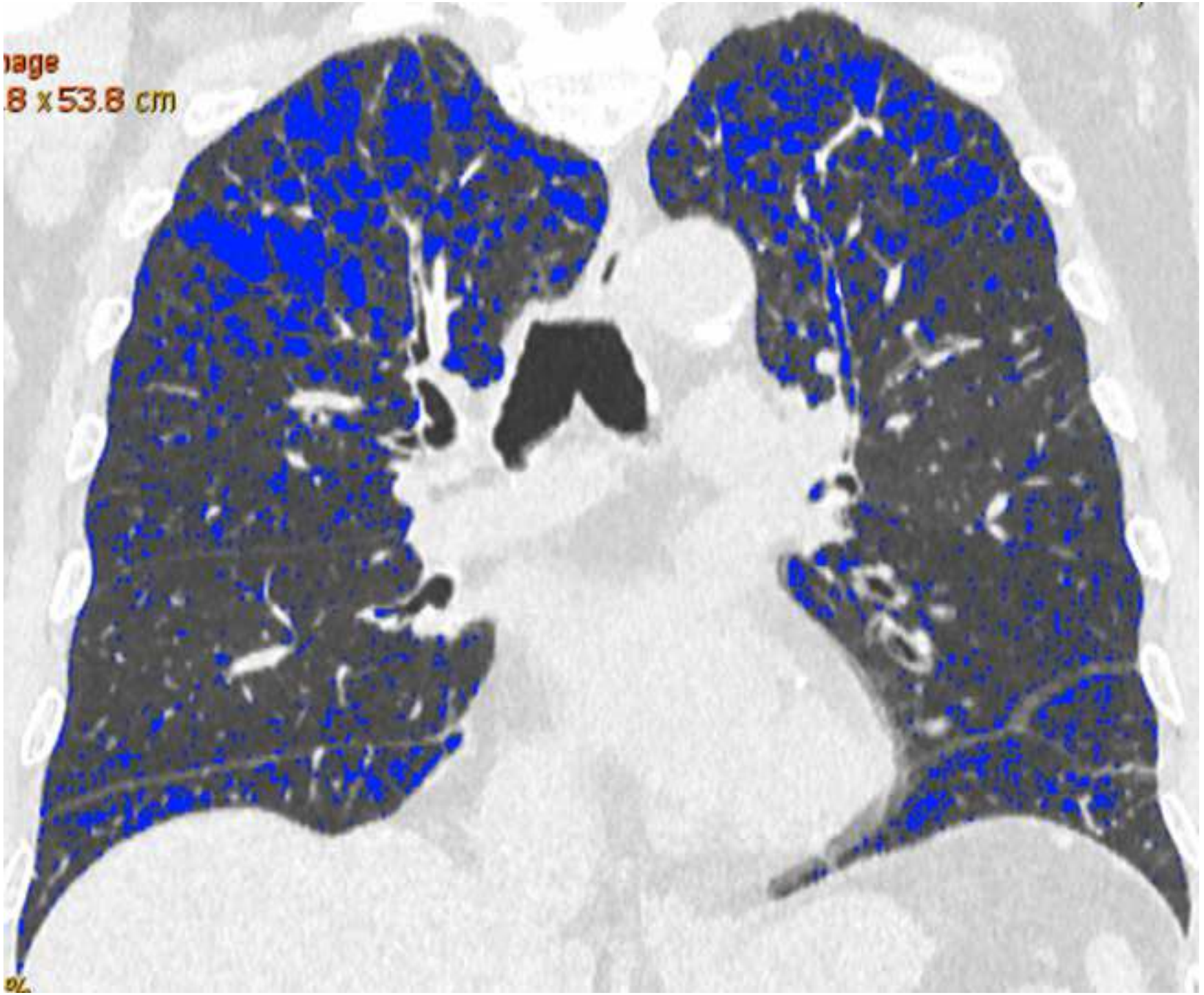

Figure 4D

Click here to access/download;Figure;Fig 4D.tif

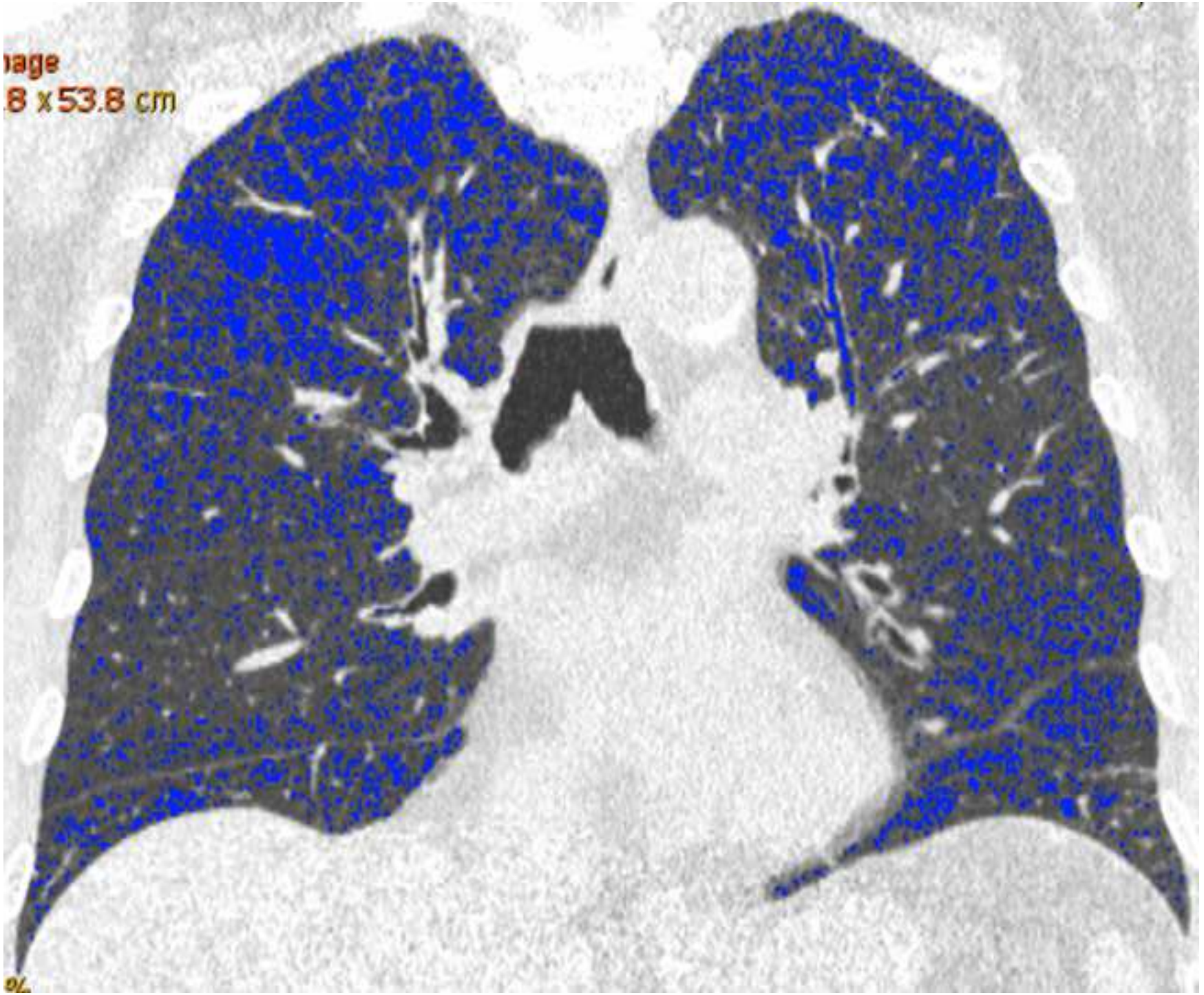

Figure 5A

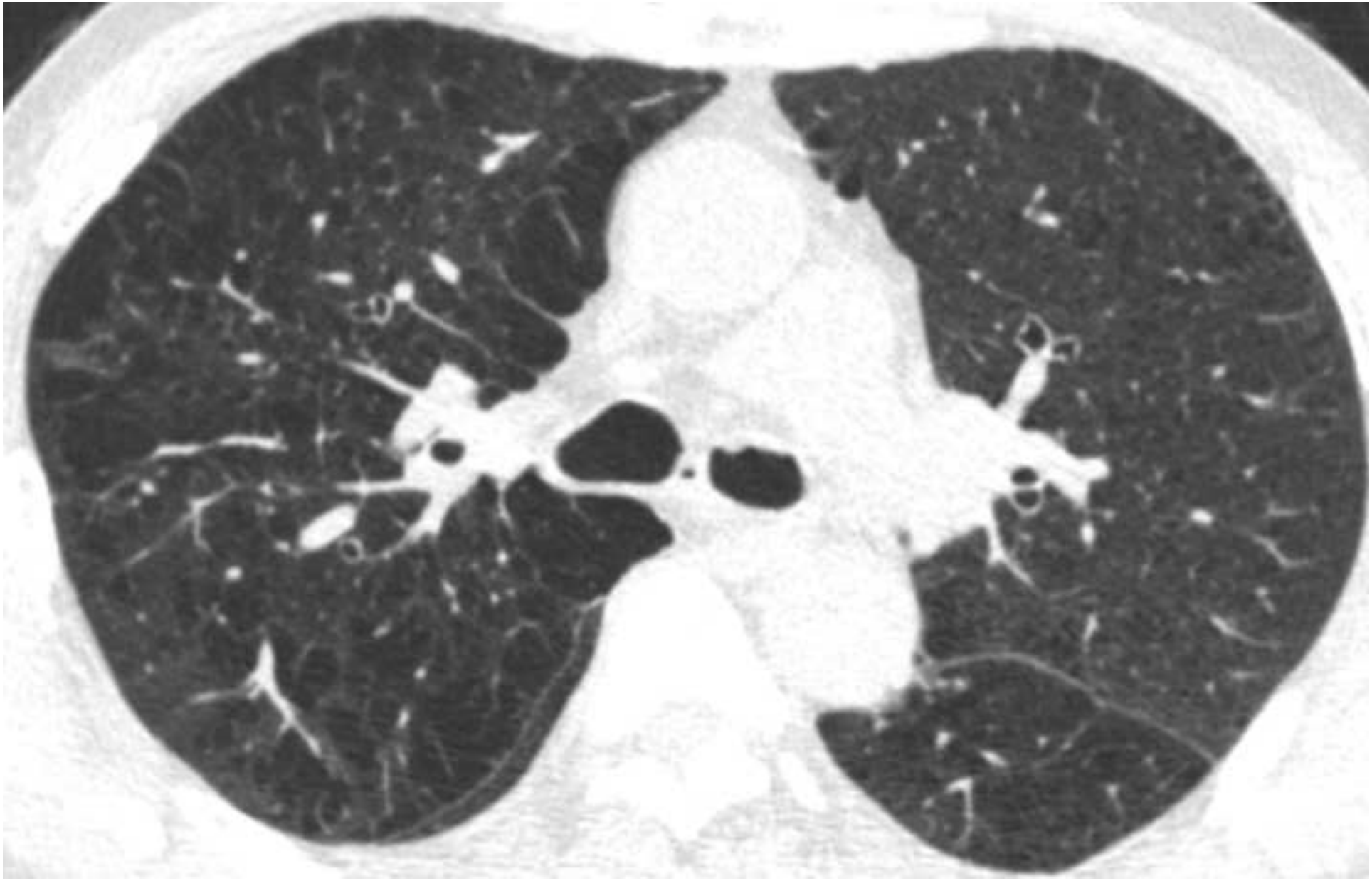

Figure 5B

Click here to access/download;Figure;Fig 5B.tif

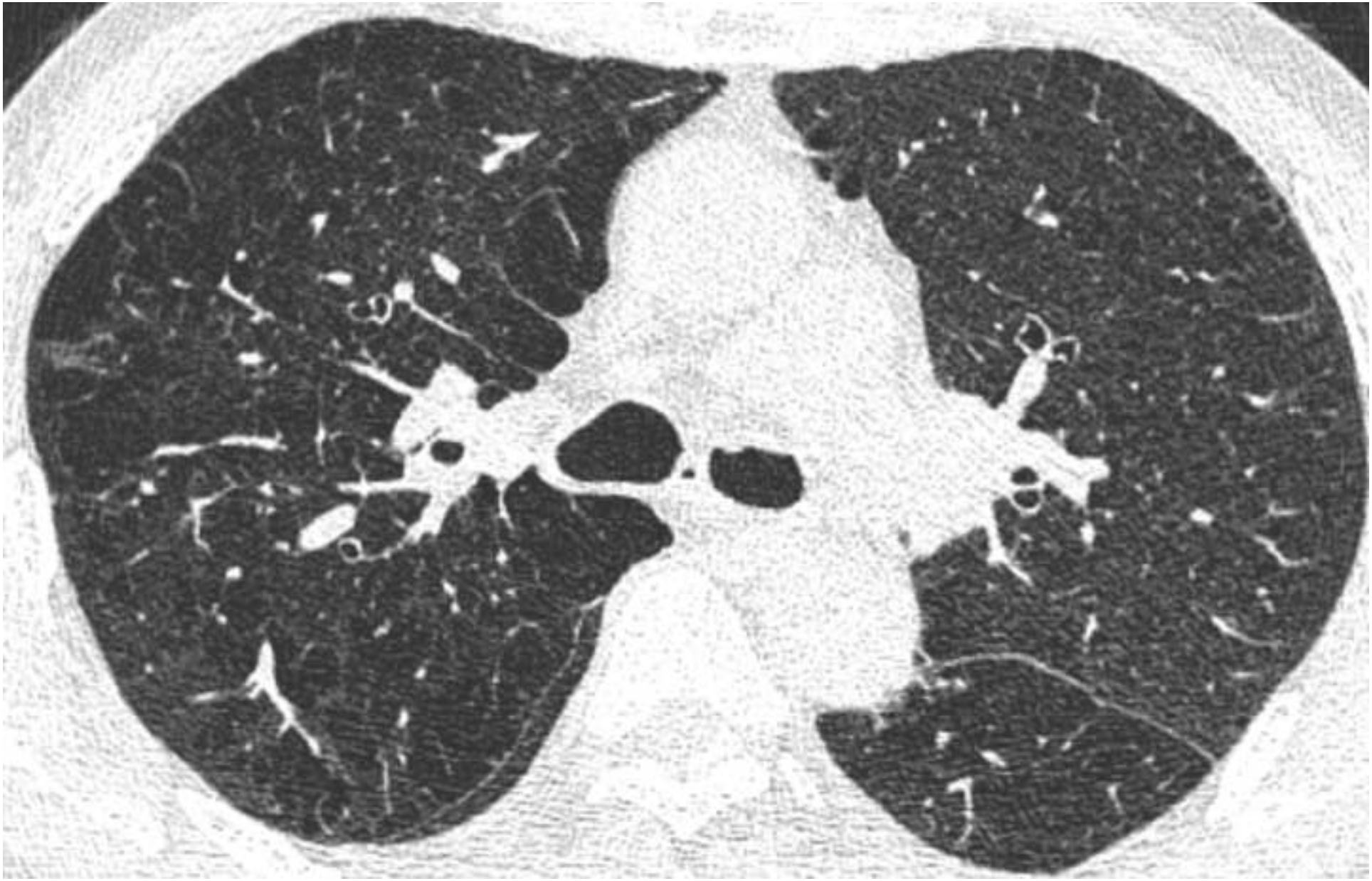

Figure 5C

Click here to access/download;Figure;Fig 5C.tif

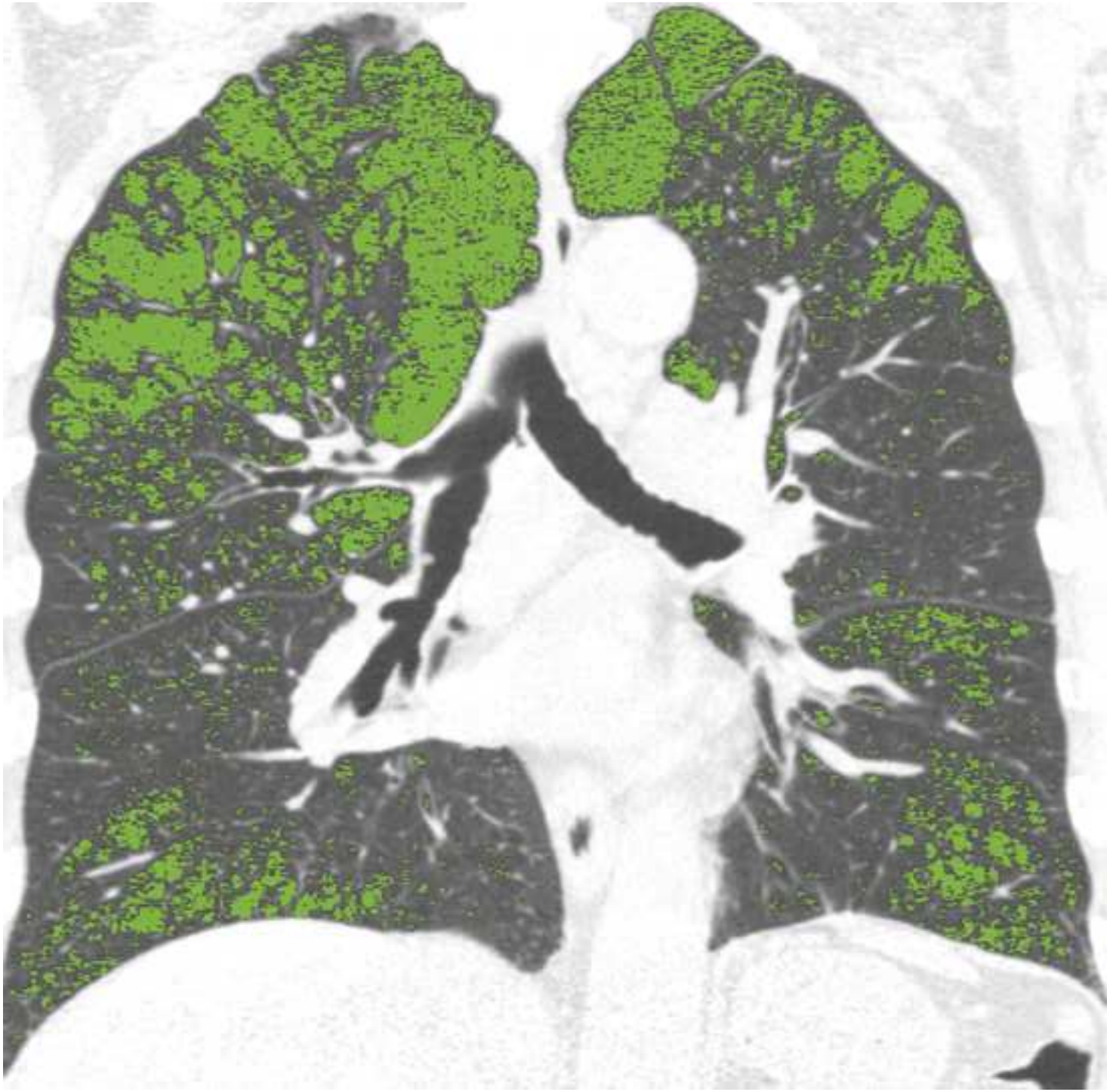

Figure 5D

Click here to access/download;Figure;Fig 5D.tif

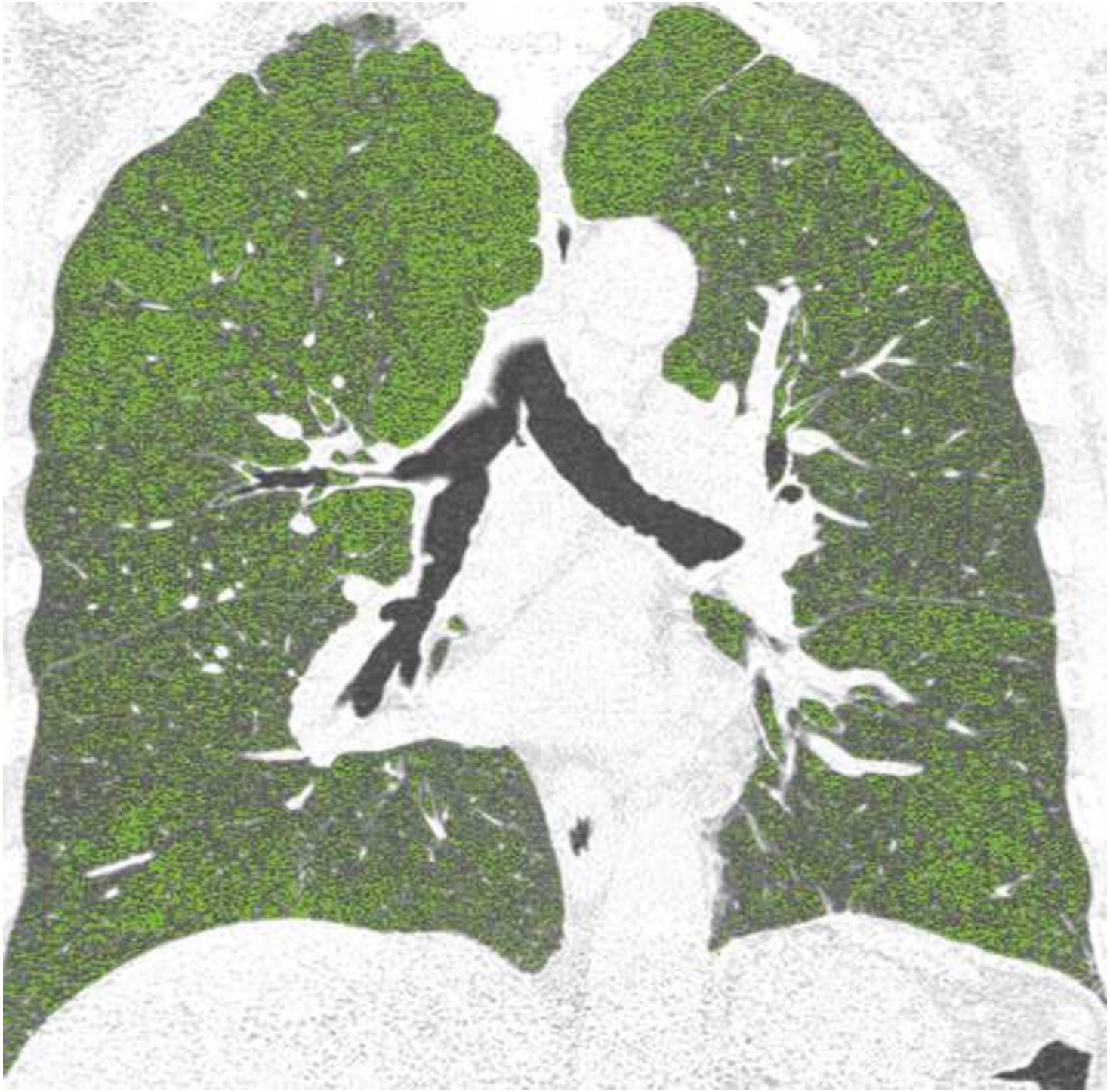

Figure 6A

Click here to access/download;Figure;Fig 6A.tif

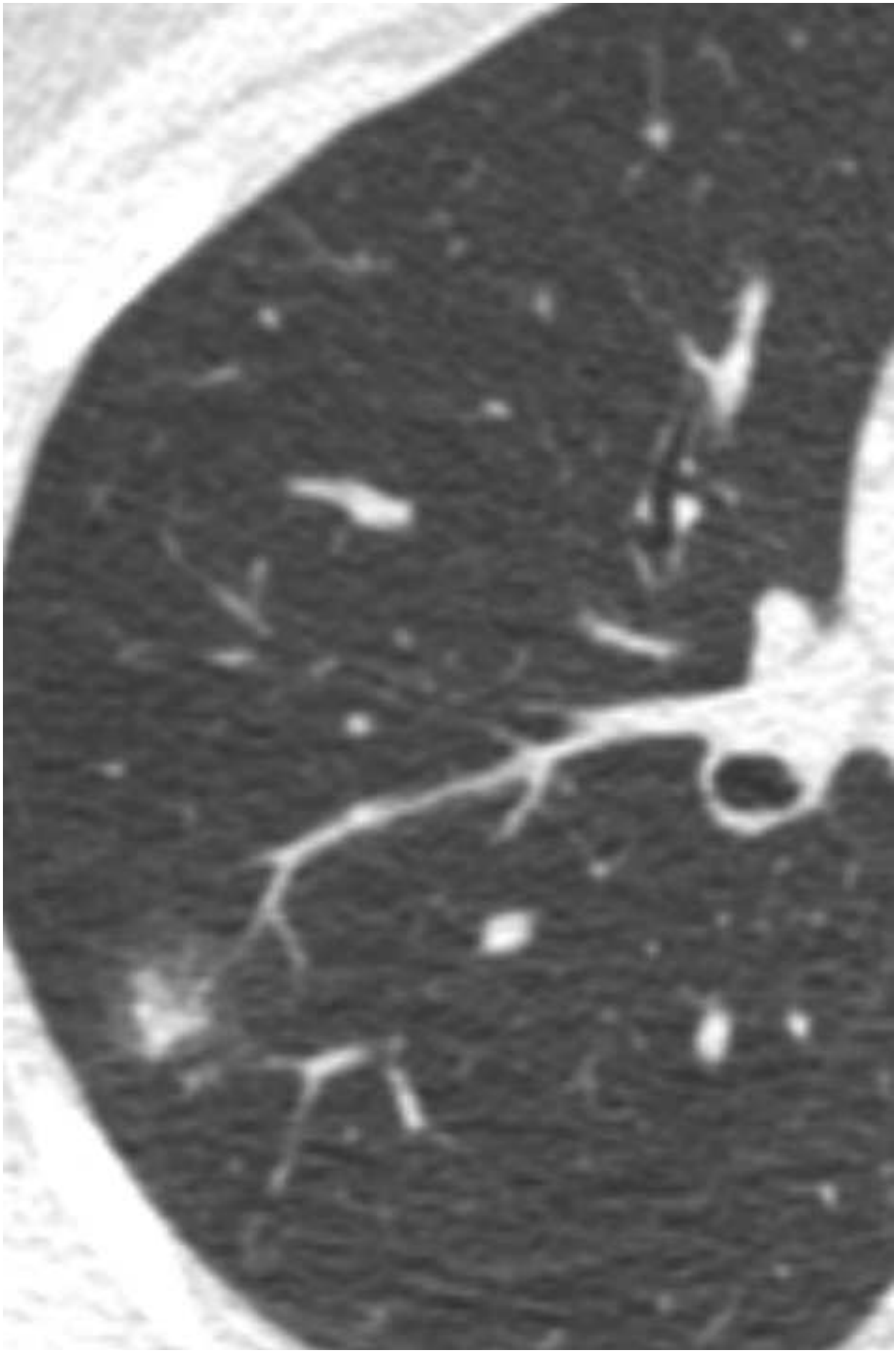

Figure 6B

Click here to access/download;Figure;Fig 6B.tif

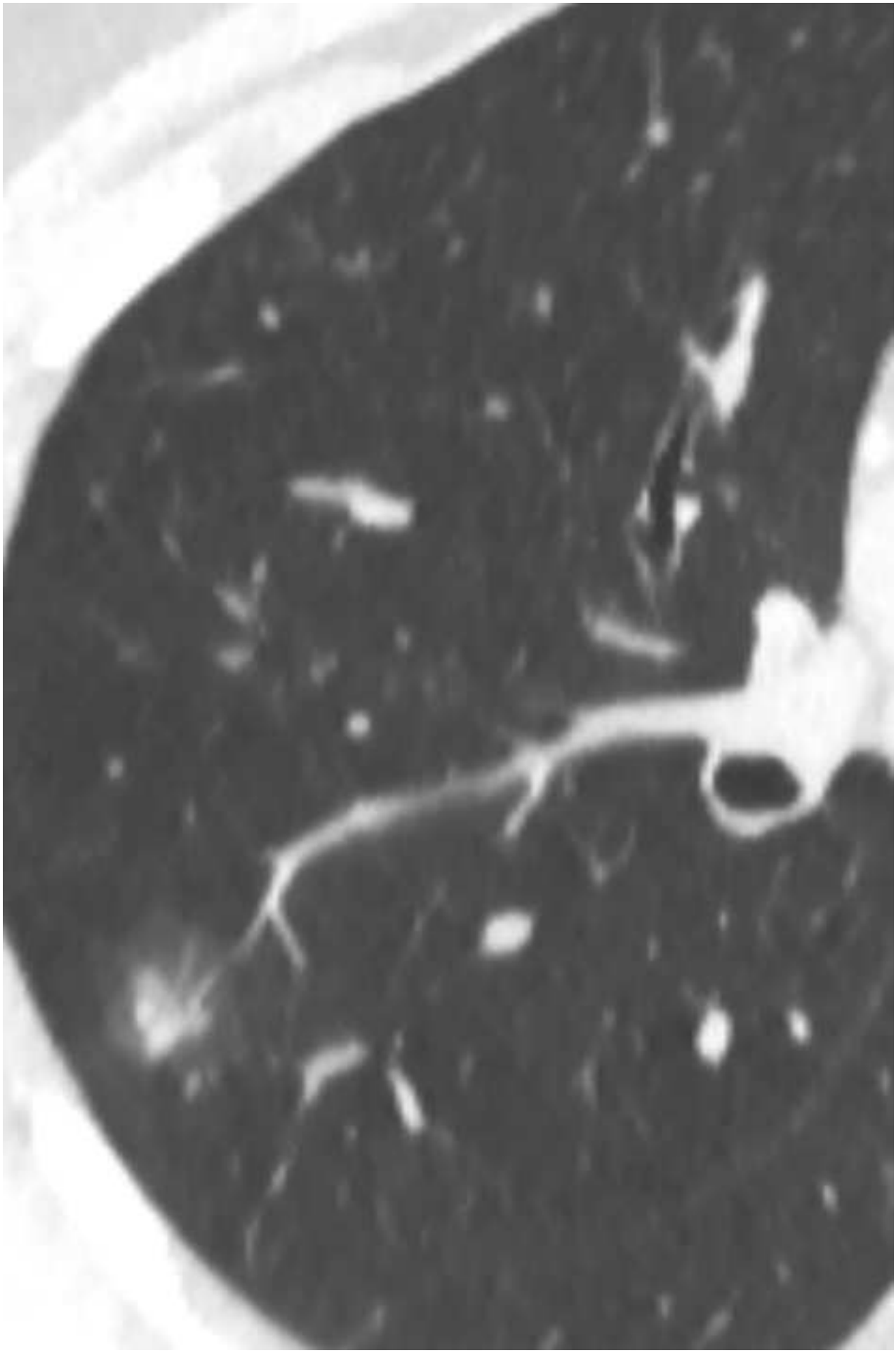

Figure 6C

Click here to access/download;Figure;Fig 6C.tif

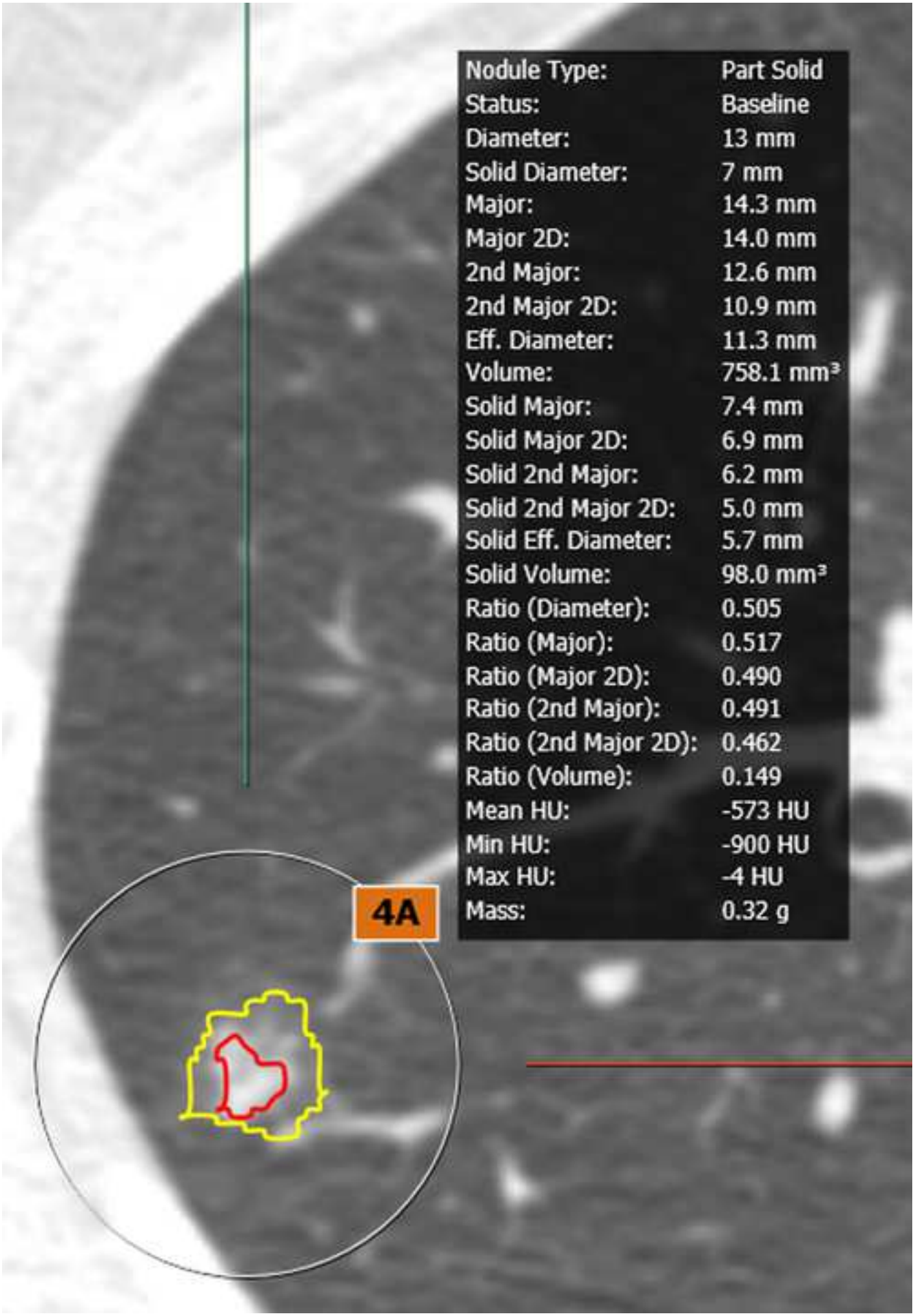

Figure 6D

Click here to access/download;Figure;Fig 6D.tif

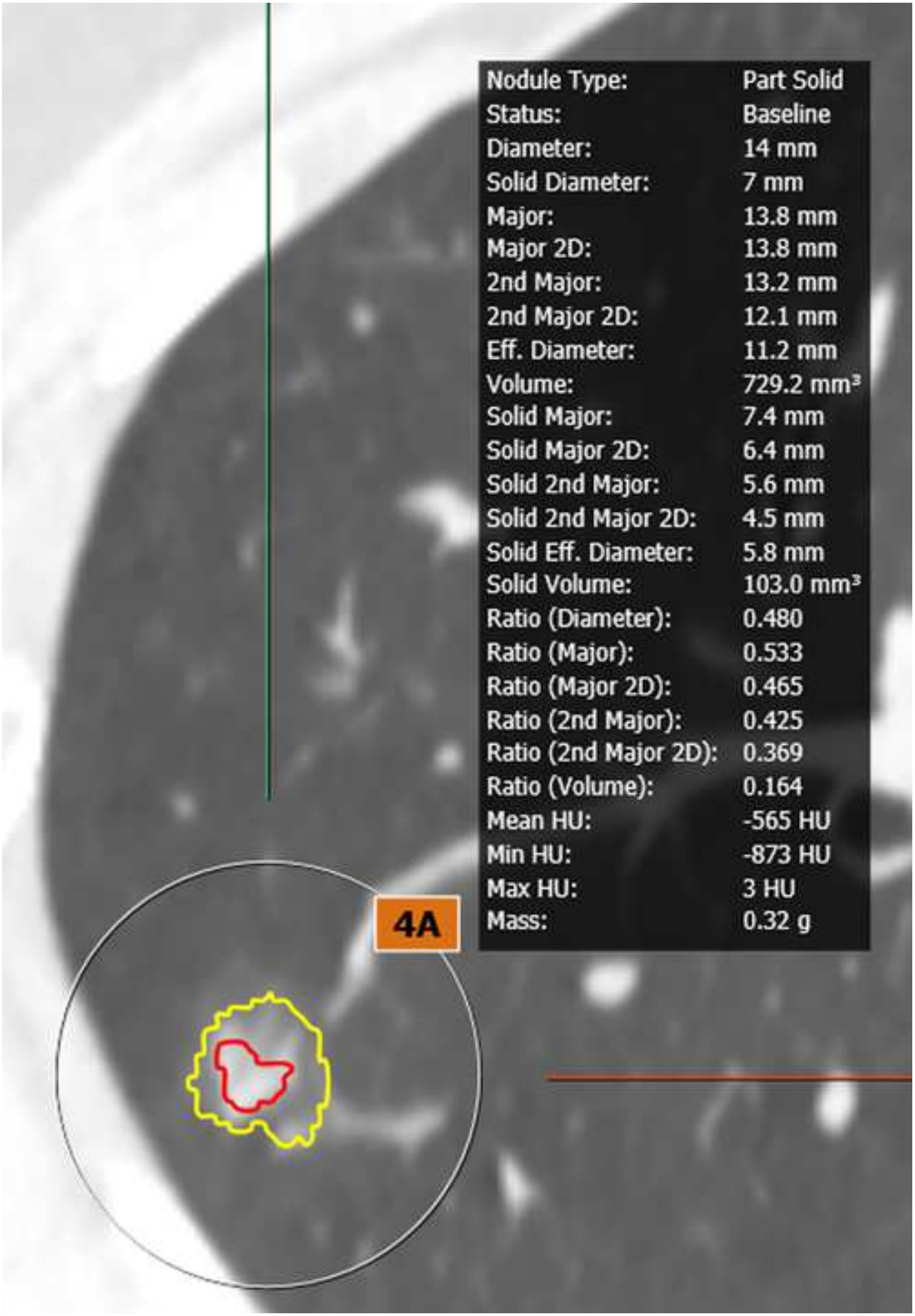

figure reprint contract