HHS Public Access

Author manuscript

Invest Radiol. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

Invest Radiol. 2015 November ; 50(11): 757–765. doi:10.1097/RLI.0000000000000180.

Measuring CT scanner variability of radiomics features

Dennis Mackin, PhD[1], **Xenia Fave, BS**[1,2], **Lifei Zhang, PhD**[1], **David Fried, BS**[1,2], **Jinzhong Yang, PhD**[1], **Brian Taylor, PhD**[3,4], **Edgardo Rodriguez-Rivera, MS**[5], **Cristina Dodge, PhD**[6], **A. Kyle Jones, PhD**[7], and **Laurence Court, PhD**[1,7]

[1]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[2]Graduate School of Biomedical Sciences, The University of Texas Health Science Center at Houston, Houston, TX 77030

[3]Research Service Line and Diagnostic & Therapeutic Care Line, Michael E. DeBakey VA Medical Center, Houston, TX 77030

[4]Department of Radiology, Baylor College of Medicine, Houston, TX 77030

[5]Radiation Oncology Department, Houston Methodist Hospital, Houston, TX 77030

[6]Department of Diagnostic Imaging, Texas Children's Hospital, Houston, TX 77030

[7]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract

Objectives—The purpose of this study was to determine the significance of inter-scanner variability in CT image radiomics studies.

Materials and Methods—We compared the radiomics features calculated for non-small cell lung cancer (NSCLC) tumors from 20 patients with those calculated for 17 scans of a specially designed radiomics phantom. The phantom comprised 10 cartridges, each filled with different materials to produce a wide range of radiomics feature values. The scans were acquired using General Electric, Philips, Siemens, and Toshiba scanners from four medical centers using their routine thoracic imaging protocol. The radiomics feature studied included the mean and standard deviations of the CT numbers as well as textures derived from the neighborhood gray-tone difference matrix. To quantify the significance of the inter-scanner variability, we introduced the metric feature noise. To look for patterns in the scans, we performed hierarchical clustering for each cartridge.

Results—The mean CT numbers for the 17 CT scans of the phantom cartridges spanned from -864 to 652 Hounsfield units compared with a span of -186 to 35 Hounsfield units for the CT scans of the NSCLC tumors, showing that the phantom's dynamic range includes that of the tumors. The inter-scanner variability of the feature values depended on both the cartridge material

Corresponding author address: Dennis Mackin, Dept. Radiation Physics, MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4017, Phone: 713-745-1642, Fax: 713-563-2545.

Conflicts of interest: none

and the feature, and the variability was large relative to the inter-patient variability in the NSCLC tumors for some features. The feature inter-scanner noise was greatest for busyness and least for texture strength. Hierarchical clustering produced different clusters of the phantom scans for each cartridge, although there was some consistent clustering by scanner manufacturer.

Conclusions—The variability in the values of radiomics features calculated on CT images from different CT scanners can be comparable to the variability in these features found in CT images of NSCLC tumors. These inter-scanner differences should be considered, and their effects should be minimized in future radiomics studies.

Keywords

radiomics; image texture; image features; CT; computed tomography; phantom

INTRODUCTION

Radiomics, the process of extracting quantifiable features from medical images, promises to improve the staging of disease and the personalization of care in radiation oncology (1, 2). For example, studies have demonstrated that extracted features can increase the stratification in Kaplan-Meier survival analysis (3, 4) and partially reproduce the global gene expression profile (5). However, to maximize the relevance and effectiveness of radiomics, the quantitative features should be reproducible and robust versus minor variations in the image acquisition parameters.

To date, most studies of the robustness of radiomics features have been based images from patients. Leijenaar *et al*. found good, but not great, agreement for 11 test/re-test patients and 23 inter-observer patients. The interclass correlation coefficients were 71% and 91% for the test/re-test and inter-observer patients, respectively(6). In another test/re-test study, Hunter *et al*. identified machine-robust image features using the concordance correlation coefficient (CCC) for test/re-test CT scans of 56 non-small cell lung cancer (NSCLC) patients(7). Balagurunathan *et al*. looked at 32 NSCLC, non-enhanced lung CT scans, and from a set of 329 features, found 29 to be repeatable (CCC $\geq$0.9) and non-redundant. Of the 29, only one was significant in survival group analysis (8).

A drawback of these coffee break-style studies is that scans are generally repeated on the same scanner after a short period of time, perhaps 20 minutes. These studies do not evaluate inter-scanner dependence or the impact of acquisition parameters, the effects of which could vary widely. For example, Kumar *et al*.(9) found large variability in the CT axial slice thickness and the pixel spacing used to acquire the 74 CT images used in a radiomics study from Basu *et al*. (10).

In general, the patient-based approach to the quality assurance of radiomics features has several problems: patient anatomy changes over time and even from one scan to the next (unless the time-scale is short); repeating scans subjects patients to additional discomfort and radiation dose; and patients are generally scanned using a single CT scanner and at one medical center making comparison of the results from different scanners/centers difficult. To address these problems, we created the Credence Cartridge Radiomics (CCR) Phantom.

By scanning the phantom with many different scanners at several different medical centers, we can more easily test intra-scanner, inter-scanner, and multi-center variability in the CT radiomics features evaluating them for both robustness and variability. We can also determine how scanner parameters such as pixel spacing, slice thickness, and dose affect the features.

The purpose of this study was two-fold. First, we compared the variability of radiomics features calculated for 17 different CT scans of the CCR phantom to the variability of the same features calculated on CT images for a set of 20 NSCLC patients. Large variability in the features relative to the inter-patient variability in tumors would indicate that the features may not be useful for discriminating tumors. Second, we investigated the inter-scanner variability of the features calculated on the CCR phantom and looked for clustering affects due to the scanner manufacturer and CT acquisition parameters. Scans acquired from four medical institutions and 16 different CT scanners produced by four different manufacturers were included in the study.

MATERIALS AND METHODS

Radiomics phantom

To facilitate the study of the inter-scanner variability of quantitative image features, we developed the CCR phantom, Figure 1(a). The phantom comprises 10 cartridges, each 10.1×10.1×3.2 cm^3, with an acrylic case. The cartridge materials were chosen to produce a wide range of radiomics feature values when scanned, ideally spanning the range of feature values found in human tissue, particularly NSCLC tumors. The first four cartridges composed of acrylonitrile butadiene styrene (ABS) plastic and were fabricated using a MakerBot® Replicator 2 3D (MakerBot Industries, LLC Brooklyn, NY) printer. These four cartridges are filed with honeycomb patterns of ABS plastic with air-filled holes of sizes of approximately 6.0, 1.4, 1.0, and 0.9 mm, making the materials 20%, 30%, 40%, and 50% filled, respectively. A block of sycamore wood provided a natural, directional texture. Two cartridges were composed of cork, one standard and one high density. The eighth cartridge, composed of rubber from shredded tires with a proprietary bonding agent (Ecoborder, Tampa, FL), had a density of 0.93 g/cm^3 and a speckled texture. The ninth cartridge, a solid block of zp® 150 power bonded with Colorbond™ (3D Systems, Inc. Rock Hill, SC) bonding agent, had the highest average density, 1.5 g/cm^3, and a barely visible texture corresponding to the pattern the 3D printer used when it sprayed the bonding resin on the powder. The last cartridge was solid polymethyl methacrylate (acrylic) with a density of 1.1 g/cm^3 and very little texture. The textures of the CCR phantom cartridges are shown in Figure 2.

Scanning the Phantom

The images used in this study were acquired using CT scanners from The University of Texas MD Anderson Cancer Center departments of Radiation Oncology, Diagnostic Imaging, and Interventional Radiology. We also used scanners from the Michael E. DeBakey Veterans Administration Medical Center, the Methodist Hospital, and Texas Children's Hospital. Scanner manufacturers included General Electric Healthcare, Philips

Healthcare, Siemens Healthcare, and Toshiba Medical Systems. All scans were acquired using the facilities' CT chest protocol and standard image reconstruction. The parameters of the scans are listed in Table 1. The GE scans were reconstructed using the *standard* reconstruction kernel which attempts to balance noise and sharpness (11). The Philips scans used either the *B* kernel, which, like the GE *standard* kernel, attempts to balance noise and sharpness, or the slightly sharper *C* kernel (12). The Siemen's *B31s* kernel is a filtered back projection with a medium smooth kernel (13), whereas the *i70f, 2* is an iterative reconstruction with strength *2* that favors sharpness (14). All of the Toshiba scans were reconstructed using the *FC18* kernel, which attempts to balance sharpness and noise in abdomen reconstructions (15).

Features Studied

This study evaluated two classes of radiomics features. First, fundamental features of the regions of interest were studied, including the mean, median, and standard deviation of Hounsfield units (HU). Entropy, which was shown to have low variability in PET images (16) and to be predictive of hepatic metastases in contrast enhanced CT images (17), was also included in this feature class. The second class of features, developed by Amadasun and King to describe human perceptible features(18), included busyness, coarseness, complexity, contrast, and texture strength. All of these features were derived from the neighborhood gray-tone difference matrix (NGTDM) and have previously been used in radiomics research (4, 19-23).

The IBEX radiomics software package was used to calculate the radiomics features (24). For each cartridge, 16 regions of interest (ROI) each with a volume of 2 cm^3 were defined as shown in Figure 3, and the average feature values were calculated. Instead of calculating features using 3-dimensional ROIs, features were calculated for each axial CT image and the results combined, a process we have termed "2.5-D" (24). This method is an alternative to the practice of calculating the features using a selected, representative slice (25-27). All scanned phantom images were re-sampled to give an isotropic in-plane pixel spacing of 1 mm^2 prior to feature calculation. The slice thicknesses were 2 to 3 mm as listed in Table 1.

NSCLC Patients

A sample of 20 CT images sets from NSCLC patients was used in this study to gauge the variability of the feature values extracted from tumors. As this is a retrospective study, informed consent from the patients was not required. All procedures were in accordance with the Declaration of Helsinki on Ethical Issues. One patient requested that access to her or his data be restricted. The average age of the other 19 patients was 67 (range: 52 – 78 years). The average heights and weights were 170 cm (154 – 182 cm) and 72.9 kg (41.0 – 97.6 kg), respectively. The mean body mass index was 25.3 (13.1 – 33.3).

Feature Noise

The feature-noise metric, $N_{m,\mathrm{f}}$ where the subscripts m and f indicate the material and the feature, respectively, was developed to compare feature variability between phantom images and patient images. $N_{m,\mathrm{f}}$ was defined as

$$N_{m,\mathrm{f}}=10*log_{10}\frac{3\sigma_{m,\mathrm{f}}}{\sigma_{p,\mathrm{f}}} \quad (1)$$

where $\sigma_{m,\mathrm{f}}$ was the standard deviation of the feature f calculated for material m from all phantom scans and $\sigma_{p,\mathrm{f}}$ was the standard deviation of the feature f calculated on the gross tumor volumes (GTV) from the set of 20 CT scans of NSCLC patients. By construction, positive $N_{m,\mathrm{f}}$ values indicated that the inter-tumor variability was less than 3 times the inter-scanner variability, and, therefore, that the variability among the scanners would be expected to make a substantial contribution to the variability measured in patient tumors. Increasingly larger values indicate relatively more noise in the features. The choice of 3 for the scaling constant was arbitrary but analogous to a 3-sigma effect, based on the assumption that the inter-patient variability should be at least three times larger than the inter-scanner variability. This choice does not impact the conclusions.

Inter-scanner Comparison

The mean and standard deviation of the HU for each scan of the CCR phantom were compared to determine whether and how the radiomics features depended on the CT scanner and acquisition parameters. The features busyness, coarseness, contrast, entropy, texture strength, and uniformity were extracted from the scans and then normalized. The normalized value $\hat{f}_{m,i}$ for feature f, material m, and scan i was calculated as:

$$\hat{f}_{m,i}=\frac{f_{m,i}-\langle f_m\rangle}{\sigma_{m,f}} \quad (2)$$

where $\langle f_m\rangle$ was the average value of feature f for material m from all the phantom scans and $\sigma_{m,\mathrm{f}}$ was the standard deviation of the feature f calculated for material m. Patterns in the scans were investigated by performing hierarchical clustering for each cartridge using the Euclidean distances of the six normalized features (28, 29).

Results

The mean and standard deviation of the CT number for the 17 scans of the CCR phantom and the 20 scans of the NSCLC tumors are shown in Figure 4. The mean for the 10 materials from the phantom ranged from -864 to 652 HU compared with a range of -186 to 35 HU for the tumors. In general, the range of mean values was larger in the set of tumors than in the individual cartridges.

The distributions for six additional features, scaled from 1 to 100, are shown in Figure 5. The feature values depended strongly on the cartridge material. Further, the range of measured values in a phantom cartridge can be large relative to the range of values observed in patient tumors, indicating that differences observed among tumor features may result from technical differences in the scanning procedure rather than actual differences in the tumors.

The feature noise metric, a relative measure of noise from CT scanner effects on radiomics features (eq. 1), is summarized for 10 features and 10 phantom materials in Table 2. The

feature noise for each feature was averaged over the 10 materials to create an overall score. Busyness and texture strength had the highest and lowest noise magnitude, respectively. The noise depended on the material, and that dependence was different for each feature.

Figure 6(a) compares the mean HUs and HU standard deviations for all of the scans. This comparison uses the results from the rubber cartridge as this cartridge has CT number and standard deviation values most similar to those in NSCLC tumors. The rubber cartridge had a mean HU of -69 compared to mean HU of -54 for tumors. The voxel mean showed some dependence on scanner manufacturer. GE scans, labeled GE1 – GE7, had voxel means that were either slightly above or slightly below average. Philips and Siemens scans, P1 – P5 and S1 - S2, had voxel means that were below average. Toshiba scans, T1 – T3, had above average voxel means.

The CT number standard deviation measured in the rubber cartridge was affected both by density differences in the material and statistical noise in the image, shown in Figure 6(b). Although the statistical noise is expected to be lower in scans using higher radiation output, this effect may be obscured by effects from other variables affecting the feature values such as the pixel spacing and image thickness. The standard deviations for all of the Philips scans were lower than average. Patterns for the other scans were less clear. However, inter-scanner differences were large, exceeding two standard deviations for four of the Philips scans and two of the Toshiba scans.

Figure 7 compares the normalized features calculated for each of the scans. None of the scans had feature values that fell outside two standard deviations from the mean. There are some apparent groupings. For example, scans P1, P3, P4, and P5 (i.e., 4 of the 5 Philips scans) appeared to have consistently similar feature values. Similarly, values for both the Siemens scans (S1 and S2) appeared to be clustered together, as were values for two of the Toshiba scans (T1 and T2). To investigate the clustering, hierarchical clustering was performed for these 6 normalized features for each cartridge, shown in Figure 8. The clustering was different for each cartridge. However, P1, P3, P4, and P5 were closely clustered in all materials except acrylic, Figure 8(e). The two Siemens scans, S1 and S2, were always clustered closely together except for the 20% fill cartridge, Figure 8(a). Thus, there is some evidence that scanner manufacturer may affect feature values. The Toshiba scan, T3, had a much larger pixel spacing (0.98 mm) than the other Toshiba scans T1 and T2 (both 0.63 mm). This combinations of larger pixel spacing and manufacturer may explain why T3 was grouped alone in all cartridges except for acrylic and 3D printed plaster, Figures 8(e) and 8(j), respectively. Similar scan parameters such as pixel spacing (even though the pixels were resampled to 1 mm during feature calculation) may explain the apparent dependence on manufacturer.

Discussion

In this study, it has been demonstrated that the variability in radiomics features extracted from CT images of the phantom was comparable in size the variability observed in the same features extracted from CT images of NSCLC tumors. The variability observed between CT scanners implies that the quality and repeatability of radiomics studies depends strongly on

the consistency of image acquisition and reconstruction. It may be possible to improve consistency by credentialing CT scanners used in radiomics studies. One approach to credentialing could be to perform a test scan of a texture phantom such as the CCR phantom. Scanners having test results that fall outside acceptable statistical control could be adjusted and retested, or, ultimately, disqualified from the study. Future retrospective studies could verify new features using the original scans. Also, any image processing techniques used in the study could be tested for consistency using texture phantom images.

Credentialing scanners as proposed may help assure some consistency of the features in radiomics studies. However, credentialing alone is not sufficient. Maximizing the information quality and benefits of radiomics requires that causes of inter-scanner variability are understood and reduced to the extent possible. In this study the feature-noise metric was introduced to quantify inter-scanner variability relative to inter-tumor variability. The feature-noise metric depended on the feature being tested. For example, the feature noise was −14.6 dB for texture strength indicating that the inter-scanner variability is quite small relative to the variability in the tumors themselves. On the other hand, the feature noise of busyness was 14.3 dB, suggesting that any signal potentially useful for differentiating tumors will be hidden in the noise from large inter-scanner variability. Future studies should attempt to develop a correction for CT scans to reduce such noise or, alternatively, to identify the features least affected by such noise to aid in feature selection.

The National Cancer Institute (NCI) has supported initiatives to promote clinical data collection for quantitative imaging (QI) and has pointed out the need for QI standards(30). One of the goals of the NCI supported Quantitative Imaging Network (QIN) is to develop a consensus on methods to validate the use of quantitative imaging (QI) so that radiomics can become a reliable part of a decision support system in radiation oncology(2). A complementary initiative, Quantitative Imaging Biomarker Alliance (QIBA), was formed by the RSNA “to improve the value and practicality of quantitative imaging biomarkers” (31). The efforts of this phantom study of the inter-scanner variability of radiomics features are in line with these goals and are complementary to other attempts to identify robust features. Through their test/re-test studies, Balagurunathan *et al*. and Hunter *et al*. have looked at the intra-scanner and intra-patient reproducibility of features(7, 8). Parmar *et al*. showed that features derived from semi-automatic contours generated by 3D-Slicer were more reproducible than those derived from physician drawn contours (intra class correlation coefficient 0.85 vs. 0.77; p=0.0009) (32).

This study limited its focus to global features derived from the neighborhood gray-tone difference matrix and on the global mean, standard deviation, and entropy of the image-intensity matrix. This group of features was appropriate for an initial study of the CCR phantom with cartridges each filled with a single material. Numerous sets of more complex features have been suggested in the literature. Kim *et al*. review the results of studies of conventional biomarkers such as tumor diameter and volume as well as the textures energy and entropy (33). Haralick *et al*. suggested that image features relating spatial and directional characteristics should be extracted from a gray-tone spatial-dependence matrix. From this matrix, they derived 14 distinct features, and each feature can be further parameterized by direction and matrix size (34). A distinct class of features, known as Gray

Level Run Lengths, was introduced by Galloway for classifying types of terrain in aerial photographs(35). Future studies should not only look at additional features such as those derived from the co-occurrence matrix and the gray-level run length, but also consider local rather than global features values. The study of local variables may require the development of additional cartridges for the CCR phantom incorporating multiple materials to produce local variability in feature values.

A more significant limitation of this study is that the variability in the features from the phantom images could result from fundamental design differences of the scanners or be caused by differences in the acquisition parameters. We intended for this preliminary study to include scan acquisition parameter variations similar to what might be seen in patient scans. An alternative approach could have been to ensure that all the scans were acquired following a strict scanning protocol that specified parameters for each scanner make and model. This alternative approach might be able to show that variability in the phantom scans was due to characteristics inherent to the scanner. If the variability were found to be small, then the scanning protocol could serve as a baseline for future patient studies. However, this alternative study would not provide an estimate of the intra-scanner variability present in existing patient scans. Future work will evaluate the impact of different imaging protocols to study to what extent such variability can be reduced.

In conclusion, this study has demonstrated that the variability in the values of radiomics features calculated on CT images from different CT scanners can be comparable in size to the variability in these features found in CT images of NSCLC tumor. To maximize the potential of any predictive models created in radiomics research, these inter-scanner differences will have to be considered, and the effects minimized. Minimizing the effects may involve credentialing CT scanners used in radiomics studies or correcting for the parameters of the scanner during data analysis.

Acknowledgments

Funding:

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number R03CA178495. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Kelly Tharp was instrumental in designing and constructing the phantom. Amy Frederick prepared the CT data sets of the 20 NSCLC patients used in the study. The authors are grateful to Craig Martin, Luke Whittlesey, Steve Parrish, Tammy Rusk, Andrea Edison, and Biju John for their assistance in acquiring the CT scans of the phantom.

References

1. Gillies R, Anderson A, Gatenby R, Morse D. The biology underlying molecular imaging in oncology: from genome to anatome and back again. Clinical radiology. 2010; 65(7):517–21. [PubMed: 20541651]
2. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012; 48(4):441–6. [PubMed: 22257792]
3. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5

4. Fried DV, Tucker SL, Zhou S, et al. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. International Journal of Radiation Oncology* Biology* Physics. 2014; 90(4):834–42.
5. Segal E, Sirlin CB, Ooi C, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nature biotechnology. 2007; 25(6):675–80.
6. Leijenaar RT, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. Acta Oncologica. 2013; 52(7):1391–7. [PubMed: 24047337]
7. Hunter LA, Krafft S, Stingo F, et al. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. Medical physics. 2013; 40(12):121916. [PubMed: 24320527]
8. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational oncology. 2014; 7(1):72–87. [PubMed: 24772210]
9. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. Magnetic resonance imaging. 2012; 30(9):1234–48. [PubMed: 22898692]
10. Basu, S.; Hall, LO.; Goldgof, DB., et al. Developing a classifier model for lung tumors in CT-scan images. Systems, Man, and Cybernetics (SMC), 2011; IEEE International Conference on, 2011; IEEE; p. 1306-12.
11. Eldevik K, Nordhøy W, Skretting A. Relationship between sharpness and noise in CT images reconstructed with different kernels. Radiation protection dosimetry. 2010 ncq063.
12. Philips Healthcare. Brillance CT Big Bore Configuration. Instructions for Use; 2007.
13. Siemens Medical Systems. Somatom Sensation 10/16 Application Guide. Siemens Medical Systems. 2006:579.
14. Xu J, Fuld MK, Fung GS, Tsui BM. Task-based image quality evaluation of iterative reconstruction methods for low dose CT using computer simulations. Physics in medicine and biology. 2015; 60(7):2881. [PubMed: 25776521]
15. Toshiba Medical Systems Corporation. Operation Manual for Toshiba Scanner Aquilion One. Toshiba Medical Systems Corporation; 2014.
16. Cook GJ, Siddique M, Taylor BP, et al. Radiomics in PET: principles and applications. Clinical and Translational Imaging. 2014; 2(3):269–76.
17. Ganeshan B, Miles KA, Young R, Chatwin C. Hepatic entropy and uniformity: additional parameters that can potentially increase the effectiveness of contrast enhancement during abdominal CT. Clinical radiology. 2007; 62(8):761–8. [PubMed: 17604764]
18. Amadasun M, King R. Textural features corresponding to textural properties. Systems, Man and Cybernetics. IEEE Transactions on. 1989; 19(5):1264–74.
19. Christodoulou CI, Pattichis CS, Pantziaris M, Nicolaides A. Texture-based classification of atherosclerotic carotid plaques. Medical Imaging, IEEE Transactions on. 2003; 22(7):902–12.
20. Mendez AJ, Tahoces PG, Lado MJ, et al. Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms. Medical Physics. 1998; 25(6):957–64. [PubMed: 9650186]
21. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. Journal of Nuclear Medicine. 2011; 52(3):369–78. [PubMed: 21321270]
22. Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. Medical Imaging, IEEE Transactions on. 2009; 28(3):374–83.
23. Giger ML, Al-Hallaq H, Huo Z, et al. Computerized analysis of lesions in US images of the breast. Academic radiology. 1999; 6(11):665–74. [PubMed: 10894069]
24. Zhang L, Fried DV, Fave XJ, et al. IBEX: An Open Infrastructure Software Platform to Facilitate Collaborative Work in Radiomics. Medical physics. 2015; 42(3):1341–53. [PubMed: 25735289]
25. Ganeshan B, Miles KA, Young RC, Chatwin CR. Texture analysis in non-contrast enhanced CT: Impact of malignancy on texture in apparently disease-free areas of the liver. European journal of radiology. 2009; 70(1):101–10. [PubMed: 18242909]

26. Skogen K, Ganeshan B, Good C, et al. Measurements of heterogeneity in gliomas on computed tomography relationship to tumour grade. Journal of neuro-oncology. 2013; 111(2):213–9. [PubMed: 23224678]

27. Ng F, Kozarski R, Ganeshan B, Goh V. Assessment of tumor heterogeneity by CT texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis? European journal of radiology. 2013; 82(2):342–8. [PubMed: 23194641]

28. R Core Team. R: A language and environment for statistical computing. 2012

29. Maechler M, Rousseeuw P, Struyf A, et al. Cluster: cluster analysis basics and extensions. R package version. 2012; 1(2)

30. Clarke LP, Nordstrom RJ, Zhang H, et al. The Quantitative Imaging Network: NCI's Historical Perspective and Planned Goals. Translational oncology. 2014; 7(1):1–4. [PubMed: 24772201]

31. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology. 2011; 258(3): 906–14. [PubMed: 21339352]

32. Parmar C, Velazquez ER, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PloS one. 2014; 9(7):e102107. [PubMed: 25025374]

33. Kim H, Park CM, Goo JM, et al. Quantitative Computed Tomography Imaging Biomarkers in the Diagnosis and Management of Lung Cancer. Investigative radiology. 2015

34. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on. 1973; (6):610–21.

35. Galloway MM. Texture analysis using gray level run lengths. Computer graphics and image processing. 1975; 4(2):172–9.

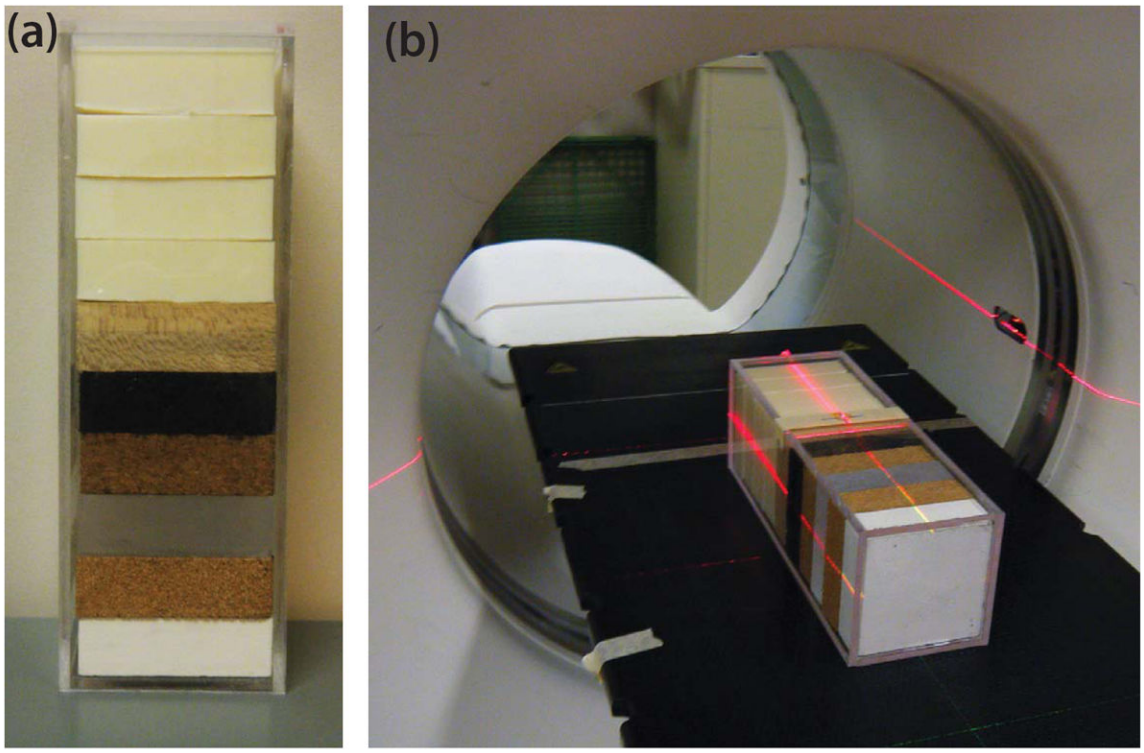

Figure 1.
(a) Credence Cartridge Radiomics (CCR) phantom with 10 cartridges. (b) CCR phantom set up for scanning.

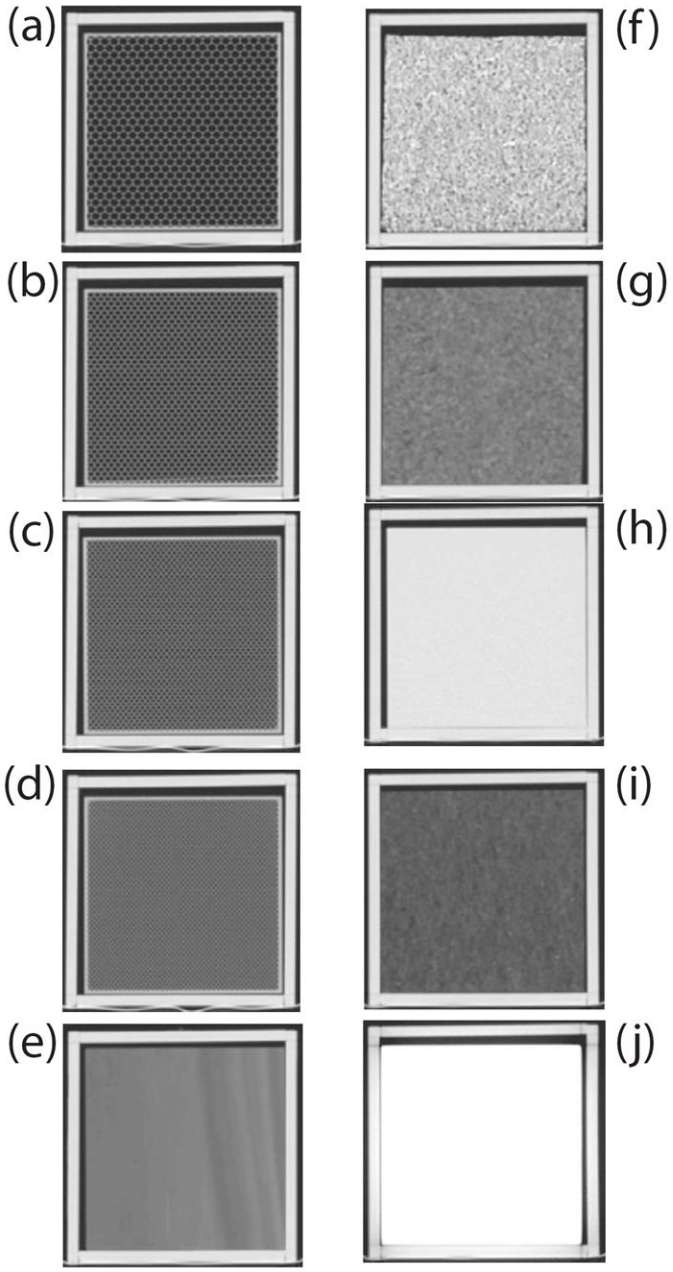

Figure 2.

Cross section of CCR phantom cartridges. (a),(b),(c),(d) are 3D printed ABS plastic with fill levels 20%, 30%, 40%, and 50% respectively. (e) is a block of natural Sycamore wood. (f) is compressed and glued rubber particles. (g) is natural cork. (h) is solid acrylic. (i) is dense cork. (j) is 3D printed solid material of a plaster based power held together with resin. Regions of higher electron density are brighter in the images. The window level is -500 HU with a width of 1600 HU.

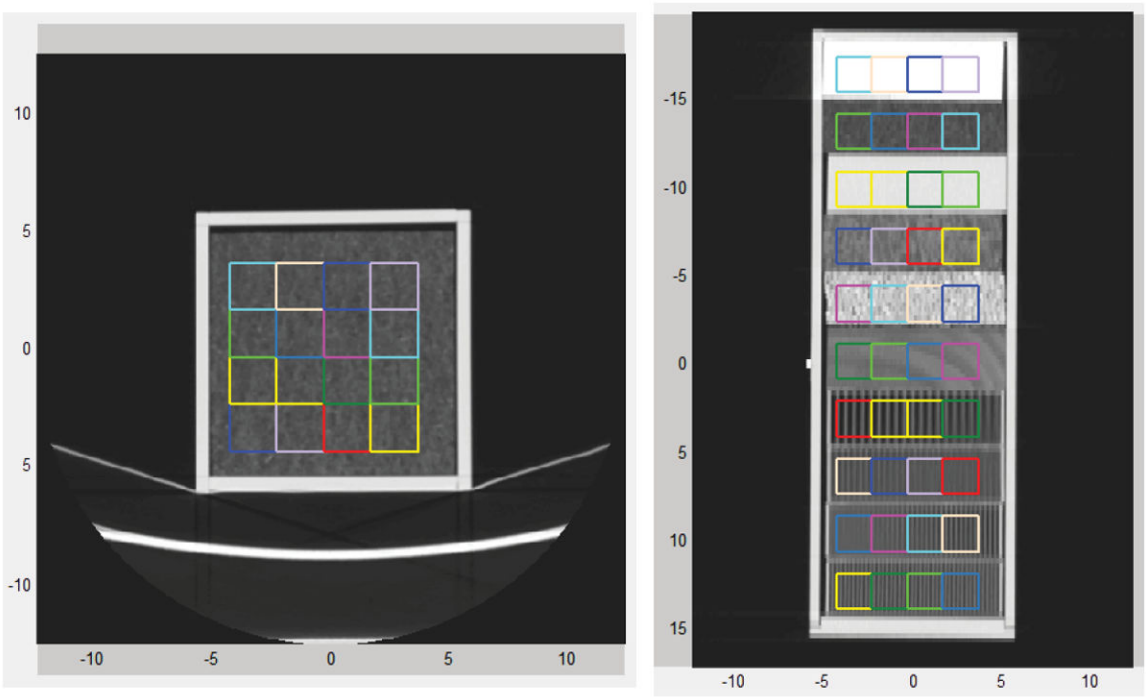

Figure 3.

Cross section (left) and coronal views (right) showing how each of the 10 layers of the phantom are divided into 16, 2 cm3 regions of interest shown as colored squares.

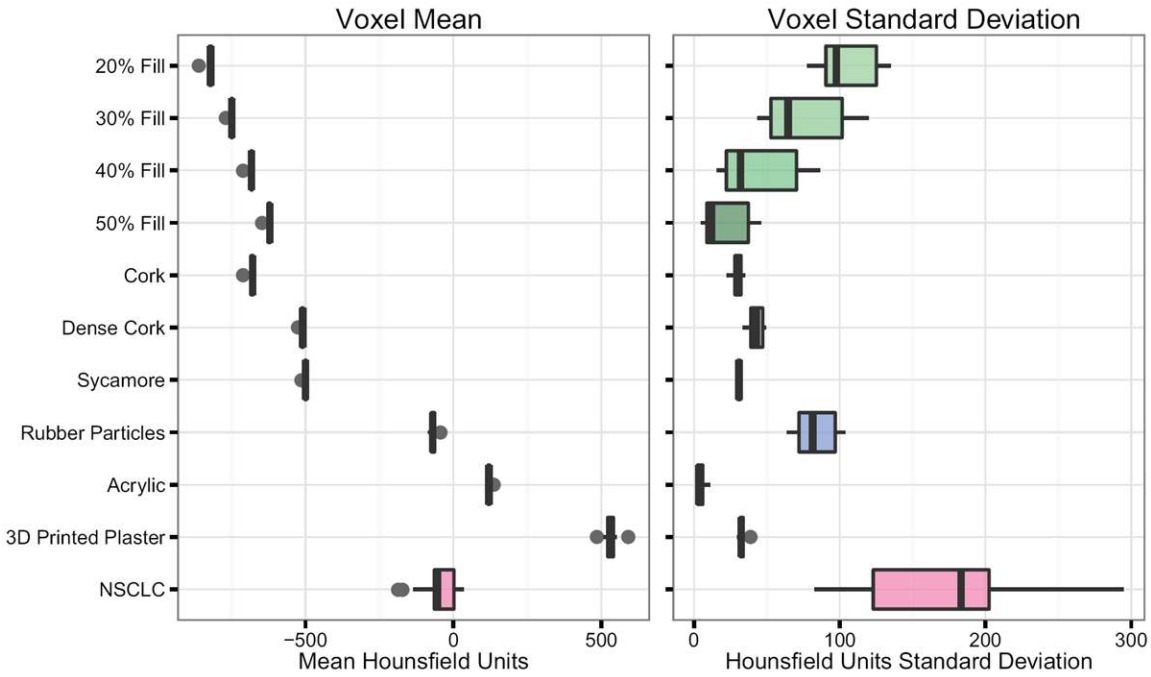

Figure 4.

The mean (left) and standard deviation (right) of the voxels for the credence cartridge radiomics (CCR) phantom and 20 non-small cell lung cancer patients.

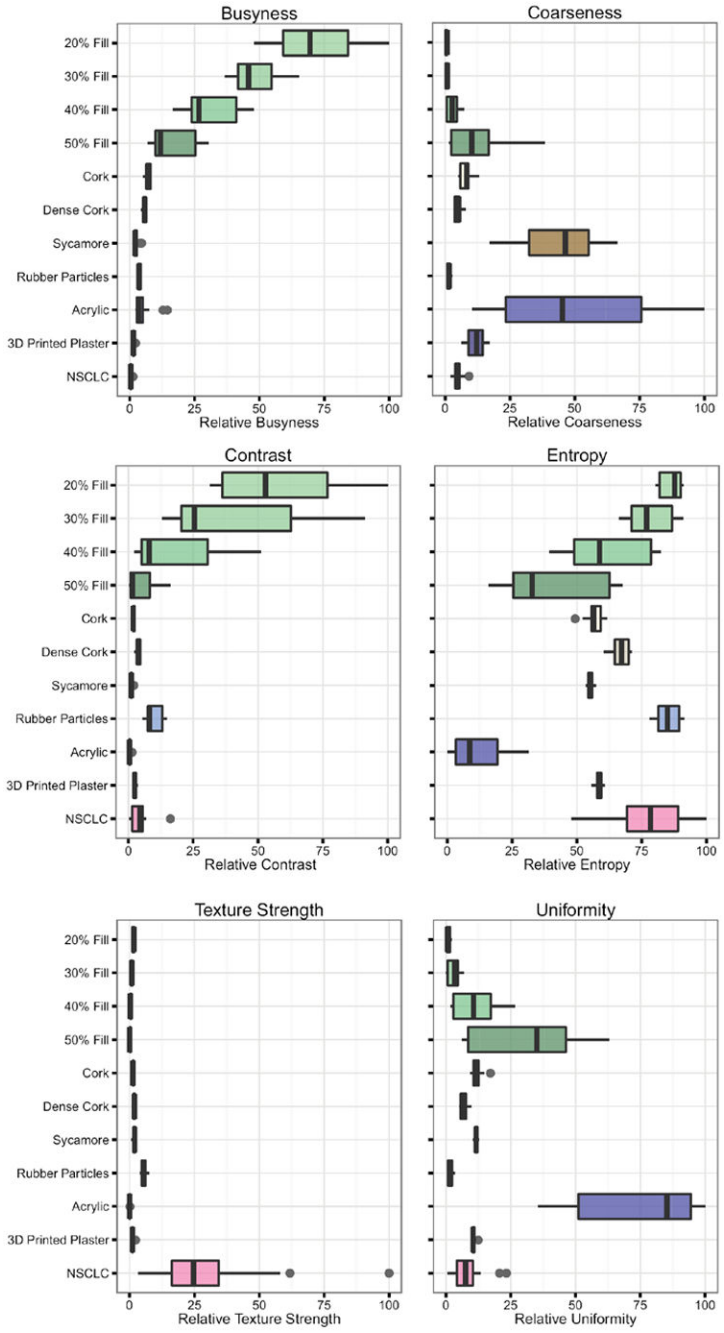

Figure 5.

Relative feature values for the credence cartridge radiomics phantom (CCR) and 20 non-small cell lung cancer tumors (NSCLC).

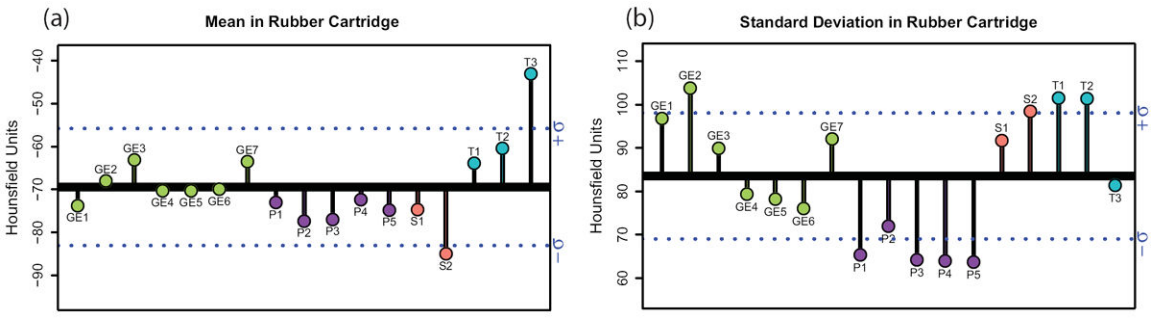

Figure 6.

The mean (left) and standard deviation (right) of the voxels for the rubber cartridge for 17 independent scans. The thick black line indicates the average value for all of the scanners. The blue dotted lines indicated one standard deviation in the measured values from all scanners. More specific information for each of the scans is provided in Table 1.

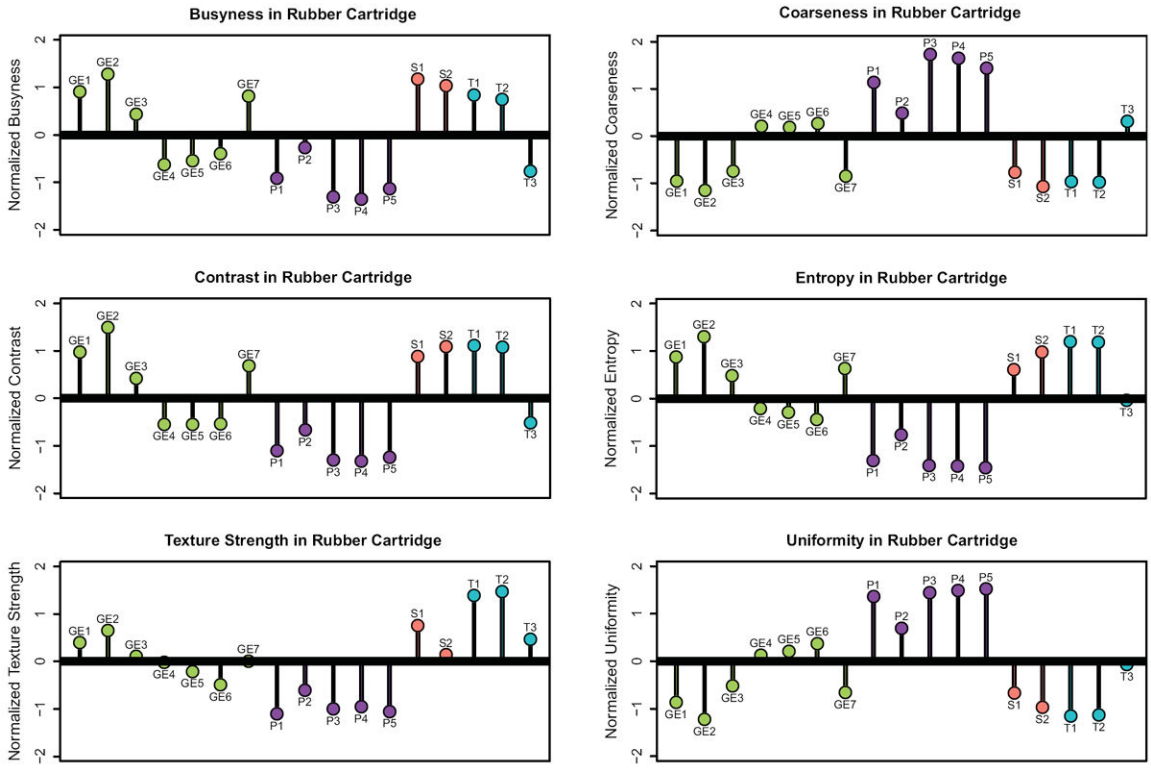

Figure 7.

The distance from the mean for six textural features calculated from the rubber cartridge of the CCR phantom. The feature values were calculated for 17 independent scans of the phantom and then normalized such that a value of 1 is equivalent to 1 standard deviation. Table 1 provides details on the scanning parameters.

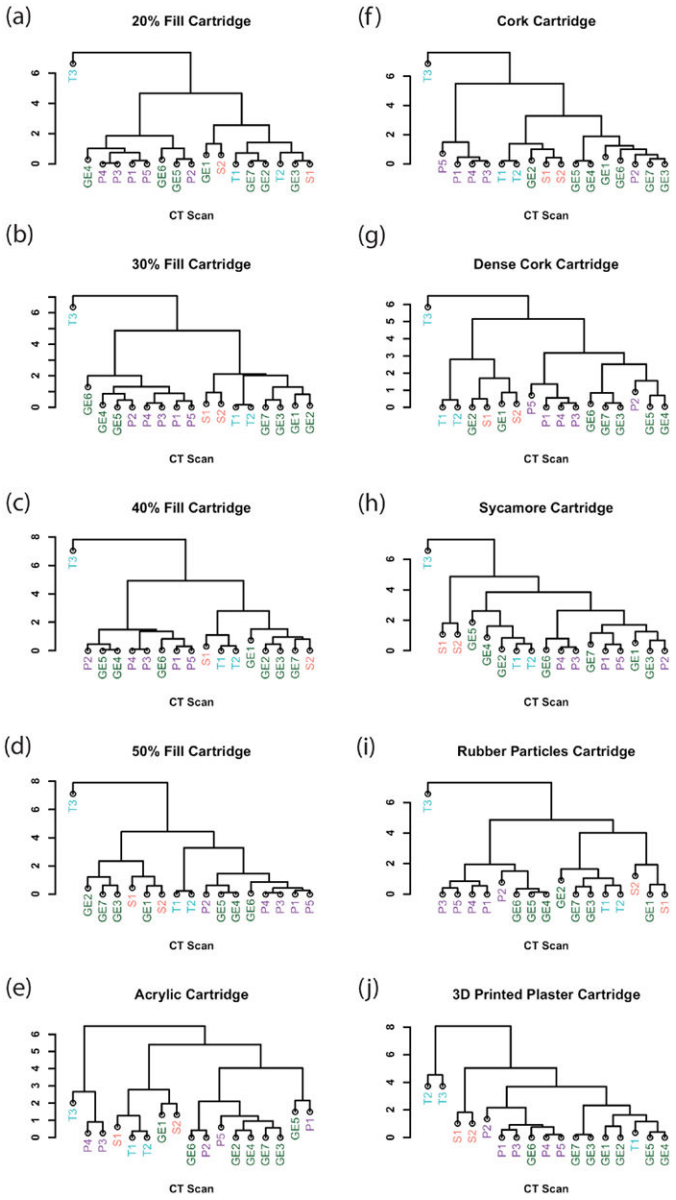

Figure 8.

Hierarchical clustering dendrogram of the 17 CCR phantom scans for each of the 10 CCR phantom cartridges: (a) 20% fill, (b) 30% fill, (c) 40% fill, (d) 50% fill, (e) acrylic, (f) cork, (g) dense cork, (h) sycamore, (i) rubber particles, (j) 3D printed plaster. The clusters were calculated from the normalized features of busyness, coarseness, contrast, entropy, texture strength, and uniformity.

Table 1

List of scanners and scan parameters used in the study. Effective mAs were calculated as mAs/spiral pitch factor. Scans S1 and S2 used variable mAs so a range of values is provided. Descriptions of the reconstruction kernels are provided in the text. Only scans GE1 and GE2 were acquired using the same CT scanner.

Scan	Manufacturer	Model	Reconstruction Kernel	Scan Type	Slice Thickness (mm)	Pixel Spacing (mm)	Spiral Pitch Factor	kVp	Effective mAs	CTDIvol (cGy)
GE1	GE	Discovery CT750 HD	standard	helical	2.5	0.49	0.98	120	81	6.19
GE2	GE	Discovery CT750 HD	standard	axial	2.5	0.70	1.00	120	300	
GE3	GE	Discovery CT750 HD	standard	helical	2.5	0.78	0.98	120	122	9.3
GE4	GE	Discovery ST	standard	helical	2.5	0.98	1.35	120	143	16.3
GE5	GE	LightSpeed RT	standard	helical	2.5	0.98	0.75	120	1102	53.6
GE6	GE	LightSpeed RT16	standard	helical	2.5	0.98	0.94	120	367	18.8
GE7	GE	LightSpeed VCT	standard	helical	2.5	0.74	0.98	120	82	
P1	Philips	Brilliance Big Bore	B	helical	3.0	0.98	0.94	120	320	17.8
P2	Philips	Brilliance Big Bore	C	helical	3.0	0.98	0.94	120	369	15.8
P3	Philips	Brilliance Big Bore	B	helical	3.0	1.04	0.81	120	320	19.9
P4	Philips	Brilliance Big Bore	B	helical	3.0	1.04	0.81	120	369	19.9
P5	Philips	Brilliance 64	B	helical	3.0	0.98	0.67	120	372	16.4
S1	Siemans	Sensation Open	B31s	axial	2.0	0.52	1.00	120	26 - 70	1.5
S2	Siemans	SOMATOM Definition Flash	I70f, 2	helical	3.0	0.54	0.60	120	17 - 28	
T1	Toshiba	Aquilion	FC18	helical	3.0	0.63	1.11	120	135	4.0
T2	Toshiba	Aquilion	FC18	helical	3.0	0.63	1.11	120	135	3.8
T3	Toshiba	Aquilion ONE	FC18	helical	3.0	0.98	0.99	120	151	13.5

Table 2

Feature noise results for 10 distinct features for each of 10 materials in the Credence Cartridge Radiomics phantom. The features are listed from most to least noisy.

	Feature Noise (dB)										
	20% Fill	**30% Fill**	**40% Fill**	**50% Fill**	**Cork**	**Dense Cork**	**Sycamore**	**Acrylic**	**Rubber**	**Resin**	**Average**
Busyness	22.8	20.1	20.8	19.7	10.4	8.9	9.7	16.1	8.6	6.0	14.3
Coarseness	-1.9	0.7	6.2	13.5	6.2	3.9	14.1	17.1	1.0	7.9	6.9
Contrast	11.8	12.4	10.2	5.9	-4.9	-2.5	-5.9	-6.6	3.5	-4.2	2.0
Median	2.0	2.1	2.3	0.9	0.7	-2.0	-1.4	-0.7	1.0	5.0	1.0
Uniformity	-4.5	0.8	6.2	10.2	-0.1	-2.1	-5.3	10.3	-2.8	-4.7	0.8
Entropy	-0.8	2.5	4.9	6.0	-1.8	-2.0	-6.9	2.8	0.0	-5.8	-0.1
Mean	-2.9	-5.6	-4.5	-4.6	-3.8	-6.5	-6.4	-5.3	-3.5	0.4	-4.3
Standard Deviation	0.2	1.2	0.8	-1.1	-7.7	-6.3	-13.0	-9.0	-1.4	-9.8	-4.6
Complexity	-0.6	-0.7	-3.7	-10.0	-13.6	-10.4	-24.9	-26.2	0.7	-17.3	-10.7
Texture Strength	-12.6	-13.8	-14.4	-16.8	-15.0	-13.6	-12.2	-25.9	-8.5	-13.4	-14.6