



Free automatic software for quality assurance of computed tomography calibration, edges and radiomics metrics reproducibility

Juan D. Saborido-Moral^{a,*}, Matías Fernández-Patón^a, Natalia Tejedor-Aguilar^b,
Andrei Cristian-Marín^c, Irene Torres-Espallardo^d, Juan M. Campayo-Esteban^c,
José Pérez-Calatayud^b, Dimos Baltas^{e,f}, Luis Martí-Bonmatí^a, Montserrat Carles^a

^a La Fe Health Research Institute, Biomedical Imaging Research Group (GIBI230-PREBI) and Imaging La Fe Node at Distributed Network for Biomedical Imaging (ReDIB) Unique Scientific and Technical Infrastructures (ICTS), 46026 Valencia, Spain

^b Department of Radiation Oncology, La Fe Polytechnic and University Hospital, Valencia, Spain

^c Department of Radiation Protection, La Fe Polytechnic and University Hospital, Valencia, Spain

^d Department of Nuclear Medicine, La Fe Polytechnic and University Hospital, Valencia, Spain

^e Division of Medical Physics, Department of Radiation Oncology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg im Breisgau, Germany

^f German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Partner Site Freiburg, Heidelberg, Germany

ARTICLE INFO

Keywords:

Computed tomography

Automatic quality assurance

Radiomics

Reproducibility

ABSTRACT

Purpose: To develop a QA procedure, easy to use, reproducible and based on open-source code, to automatically evaluate the stability of different metrics extracted from CT images: Hounsfield Unit (HU) calibration, edge characterization metrics (contrast and drop range) and radiomic features.

Methods: The QA protocol was based on electron density phantom imaging. Home-made open-source Python code was developed for the automatic computation of the metrics and their reproducibility analysis. The impact on reproducibility was evaluated for different radiation therapy protocols, and phantom positions within the field of view and systems, in terms of variability (Shapiro-Wilk test for 15 repeated measurements carried out over three days) and comparability (Bland-Altman analysis and Wilcoxon Rank Sum Test or Kendall Rank Correlation Coefficient).

Results: Regarding intrinsic variability, most metrics followed a normal distribution (88% of HU, 63% of edge parameters and 82% of radiomic features). Regarding comparability, HU and contrast were comparable in all conditions, and drop range only in the same CT scanner and phantom position. The percentages of comparable radiomic features independent of protocol, position and system were 59%, 78% and 54%, respectively. The non-significant differences in HU calibration curves obtained for two different institutions (7%) translated in comparable Gamma Index G (1 mm, 1%, >99%).

Conclusions: An automated software to assess the reproducibility of different CT metrics was successfully created and validated. A QA routine proposal is suggested.

1. Introduction

Computed Tomography (CT) imaging is a consolidated modality for the diagnosis, staging, treatment, and prevention of multiple diseases, consolidated and highly available [1,2]. Among all its possible applications, it highlights its relevance in oncology, playing a pivotal role in cancer early diagnosis and monitoring treatment effects [3,4]. CT images play a crucial role in radiotherapy planning, allowing the delineation of tumors and organs at risk. Moreover, CT is commonly used for

dose calculation [5,6], where the dose of the different organs is calculated from the electronic density obtained from the CT images.

Technical improvements have made possible to extract high-throughput quantitative features from images, known as radiomics [7], allowing data mining and analysis. Radiomic features provide information about tumor shape, microarchitecture, and heterogeneity. Radiomics are used to construct either descriptive or predictive models that facilitate clinical decision making. This is particularly relevant in oncology, where radiomics may thus give important surrogate

* Corresponding author.

E-mail address: juansamo@alumni.uv.es (J.D. Saborido-Moral).

<https://doi.org/10.1016/j.ejmp.2023.103153>

Received 20 June 2023; Received in revised form 16 September 2023; Accepted 22 September 2023

Available online 30 September 2023

1120-1797/© 2023 Associazione Italiana di Fisica Medica e Sanitaria. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

phenotypic information [8], providing significant data to determine survival and tumor response [9]. Nevertheless, radiomics features have proven to be sensitive to variations in acquisition parameters, signal to noise ratios, image processing methods and tumor delineation [9], limiting the generality of prediction models built from radiomic features.

Quality Assurance (QA) protocols are needed to assess CT systems performance, the quality of the obtained images, and the robustness and reproducibility of the different features. Periodic QA tests regarding image quality are recommended in the AAPM TG-66 report [10] and by international authorities such as ICRP and IAEA [11]. Moreover, institutions like ICRU and the American College of Radiologists (ACR) emphasize the effectiveness of quantitative tools for evaluating phantom images in QA tests [12,13]. Although most CT manufacturers provide their own commercial software for routine QA programs, they rely upon the vendor and require the use of their specific phantoms. In addition, their implemented metrics are rather elementary and far from being clinically relevant. Moreover, the closed software approach does not allow user interaction, denying the possibility of adding new features more suitable for the user requirements. Consequently, multiple open-source QA programs have been developed, which require the use of specific phantoms [14–16]. However, they do not evaluate the reproducibility of the calculated metrics, neither integrate the characterization of tissues interface borders nor radiomic features.

In this study, an open-source software solution is presented for the automation of QA in CT systems, fully developed using Python. It automates the QA process by using reference segmentations. The QA procedure calibrates the HU, characterizes tissue interface borders and extracts radiomic features using the PyRadiomics platform [17]. To assess generality, the tool was validated in six CT scanners using two different phantoms. As a proof of concept, the proposed QA software was used to evaluate the dependence of the acquisition protocols, the position of the phantom and the CT scanner on reproducibility.

2. Methods

2.1. Experimental phantoms

Two different experimental phantoms were used: the Electron Density Phantom Model 062 M (CIRS) and the Tomotherapy Cheese Phantom (Accuray). Both consist of a cylindrical container with a similar to water electronic density and different holes where inserts simulating different human tissues are placed.

2.2. CT scanners

Six different CT systems from two different institutions were used. From Center 1: Philips Gemini TF 64 PET/CT from Nuclear Medicine department (PET/CT-NM); Philips Gemini TF BigBore CT from Radiation Oncology department (CT-RT) and from Diagnostic Radiology department the Philips Brilliance iCT 256 (CT-DR-1) and the Toshiba Aquilion 64 CT (CT-DR-2). From Center 2: Philips Gemini TF PET/CT from Nuclear Medicine department (PET/CT-NM-F) and Philips Brilliance 16 CT from Radiation Oncology department (CT-RT-F).

2.3. Protocols

The different protocols evaluated in each PET/CT system can be found in the [Supplementary Material \(Table S1\)](#).

2.4. Metrics

The evaluated metrics were divided in three groups.

2.4.1. HU calibration

The software characterizes the electronic density from the measured

HU. For this purpose, the reference segmentations are placed in the middle of each one of the inserts of the phantom (Fig. 1a) and a calibration curve relating the physical density and the HU is calculated.

2.4.2. Edges characterization

The reference segmentations cover completely the insert up to the edge (Fig. 1b). The software uses contrast and drop range to characterize the edges. Contrast metric evaluates the intensity difference between the insert and the phantom body calculating an intensity gradient in all the voxels that make up the edge region. Drop Range characterizes how steep the intensity drop is on the edge of the insert. For this purpose, the pixel intensities along four directions on the transversal plane (covering the interface between the insert and the phantom body) are computed to create an intensity profile. Then, an interval was defined taking the pixels where the intensity values laid between the 10% and the 90% of the maximum intensity value of the intensity profile. Thinner intervals represent more defined borders and higher intensity drops.

2.4.3. Radiomic features

The reference segmentations defined for the computation of radiomic features were nine inhomogeneous areas with different electronic densities and two homogeneous areas in air and water (Fig. 1c). A total of 45 different metrics were calculated, classified as First Order metrics, Gray Level Co-occurrence Matrix metrics (GLCM), Gray Level Size Zone Matrix metrics (GLSZM), Gray Level Run Length Matrix metrics (GLRLM) and Neighboring Gray Tone Difference Matrix (NGTDM) radiomic features.

2.5. Automatic quality assurance workflow

The software uses a reference image of the phantom to automate the QA process. The tool resizes and rigidly registers the reference image to the new image, saving the result in a transformation matrix. The transformation matrix is subsequently used to transform the reference segmentations to fit to the new images and are then used to calculate the metrics. (Fig. 2).

2.6. Reproducibility analysis

2.6.1. Variability

For the intrinsic variability, the CIRS phantom was imaged with the PET/CT-NM system from Center 1 with the protocol commonly used in clinical practice (protocol C from [Table S1](#)). The phantom was imaged 5 times per day, 3 different days. The variability of each metric was studied over the 15 acquisitions, evaluating the goodness-of-fit of the data distribution to a normal gaussian distribution, by using the Shapiro-Wilk normality test [18].

2.6.2. Comparability

The comparability of the metrics was analyzed in terms of the implemented protocol, the position of the phantom inside the field of view (FoV) and the CT system used. The protocol C from PET/CT-NM ([Table S1](#)) was used as reference protocol. An overview of the assessment of the comparability of the metrics is shown in [Fig. 3](#).

A Bland-Altman analysis [19] was carried out to assess the comparability of the HU calibration. For edge characterization metrics, Kendall Rank Correlation Coefficient (KRCC) [20] was implemented to evaluate the similarity between two ordinal classifications. Finally, both in HU calibration and radiomic features characterization a Wilcoxon Rank Sum Test (WRST) [21] was also carried out.

2.6.3. Gamma index

To assess the effect that the different calibration curves may have in radiotherapy planning the gamma index [22] was calculated for four different radiotherapy plannings in different cancer sites: lung, brain, prostate and head and neck areas. The planning results using the

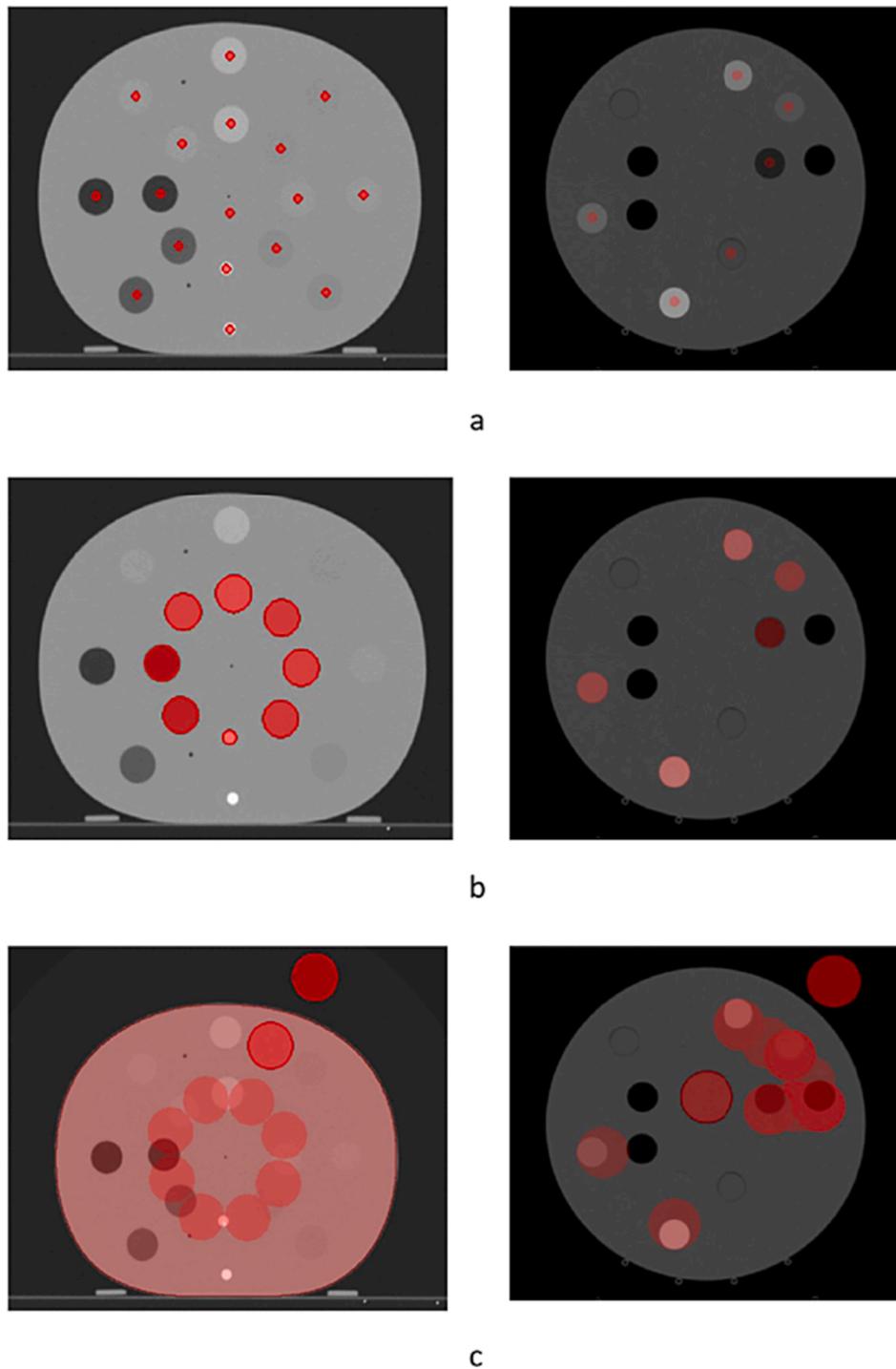


Fig. 1. Reference segmentations for the computation of (a) HU calibration, (b) edges characterization and (c) radiomic features in CIRS phantom (left) and Cheese phantom (right).

calibration curve from NM department at Center 1 were compared to the ones obtained by implementing the curve from RT department at Center 1 and the curve obtained from Center 2. The dose difference criterion was set to a 1 % and the distance-to-agreement (DTA) to 1 mm.

3. Results

Our open-source code for automatic CT QA can be downloaded from https://github.com/juandasm/CT_Metrics_Reproducibility.

3.1. Intrinsic variability of the metrics

For the segmentations shown in Fig. 1, the code was employed to evaluate if the values of the HU, edge parameters and radiomics features followed a normal distribution across the 15 acquisitions. All acquisitions were performed with the same protocol (C), same CT scanner (PET/CT-NM at Center 1) and same position of the CIRS phantom (center of FoV). Results are shown in Table 1. In 15 out of the 17 inserts, the HU values followed a normal distribution. For edge parameters, both contrast and drop range followed a normal distribution in 5 out of the 8 inserts. From the contrast and drop range values a classification of the

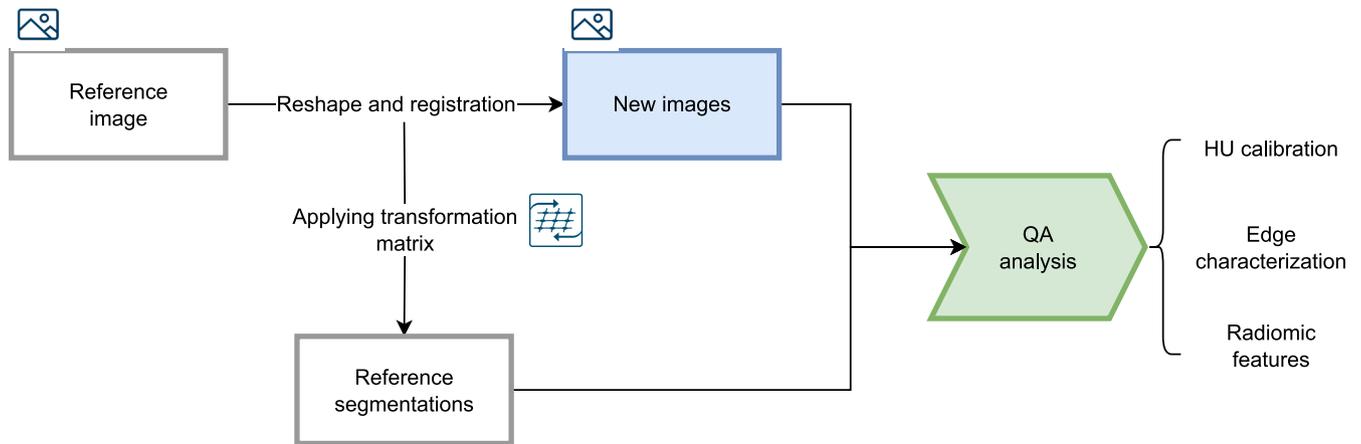


Fig. 2. Automatic QA software workflow.

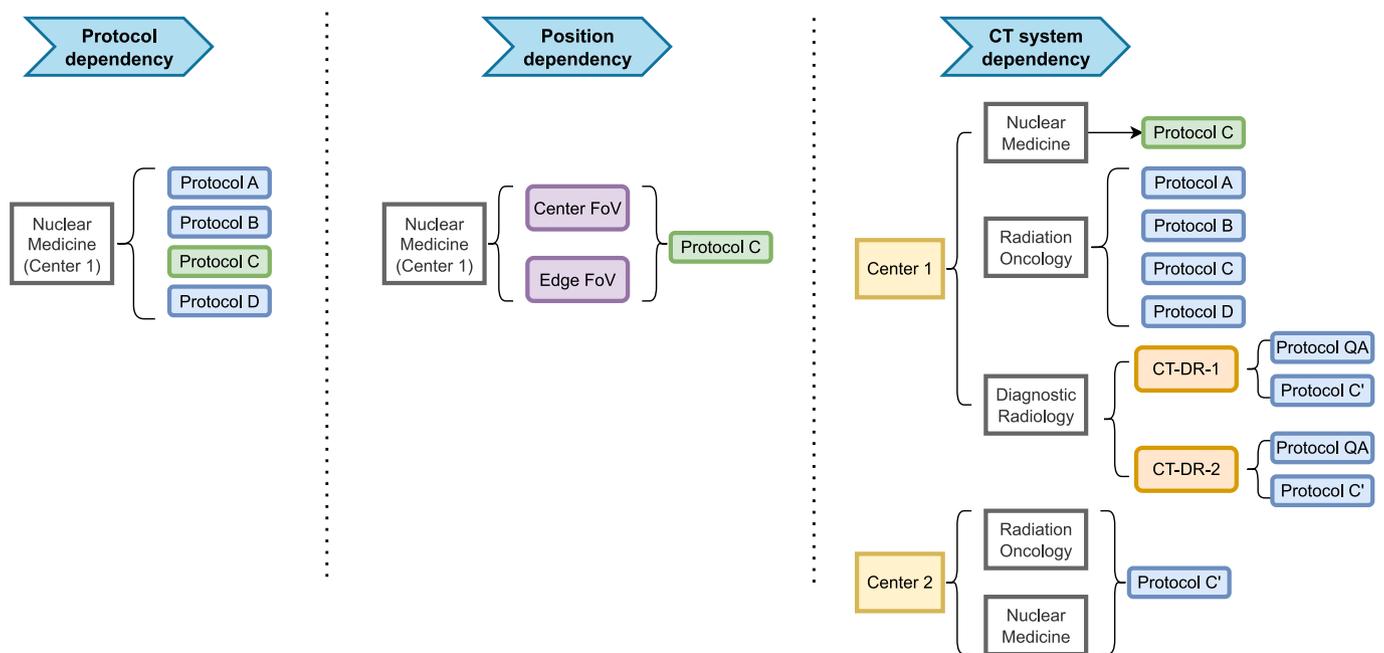


Fig. 3. Diagram showing the comparability analysis divided in dependency on protocol, position, and CT system.

Table 1

Mean value and standard deviation of HU, contrast and drop range for each tissue density. Values in bold if the measurements did not fit to a normal distribution in all the inserts and in white if they did. Measurements carried out with PET/CT-NM from Center 1 implementing protocol C.

Insert	HU	Contrast	Drop range
Trabecular Bone	216 ± 4	840 ± 31	0.57 ± 0.09
Breast	-41 ± 3	116 ± 14	0.49 ± 0.18
Muscle	30 ± 5	146 ± 14	0.46 ± 0.12
Adipose	-75 ± 2	141 ± 14	0.49 ± 0.10
Dense Bone	827 ± 9	10394 ± 1625	0.67 ± 0.15
Lung (Exhaling)	-487 ± 2	2845 ± 124	0.52 ± 0.08
Lung (Inhaling)	-787 ± 4	7808 ± 280	0.62 ± 0.11
Liver	41 ± 3	170 ± 29	0.49 ± 0.15
Water	-18 ± 3		

inserts was obtained for each measurement; the most repeated classifications are presented in Fig. 4. When comparing classifications by KRCC, both contrast and drop range were comparable. Therefore, classification instead of absolute values is employed in the following sections. Finally,

for the 45 radiomic features evaluated, 37 showed a normal distribution in at least 8 of the 11 segmentations. Only these 37 radiomic features will be evaluated in the following sections.

3.2. Comparability of metrics

All results are summarized in Table 2.

3.2.1. Protocol dependency

Impact of protocol was evaluated with the PET/CT-NM system at NM department in Center 1 and with the CIRS phantom placed at the center of the FoV. The results derived from protocol C were compared with the other three protocols for the PET/CT-NM system (Table S1). Analyzing the measurement of the HU, both the BA analysis, and the WRST confirmed that all protocols were comparable. Based on these results, a recommended calibration curve was calculated averaging across all protocols of NM. Calibration curves for each protocol and the recommended calibration curve are shown in Fig. 5. The HU calculated with this recommended calibration curve differed in less than a 5% with respect to the HU calculated with the calibration curve obtained for each

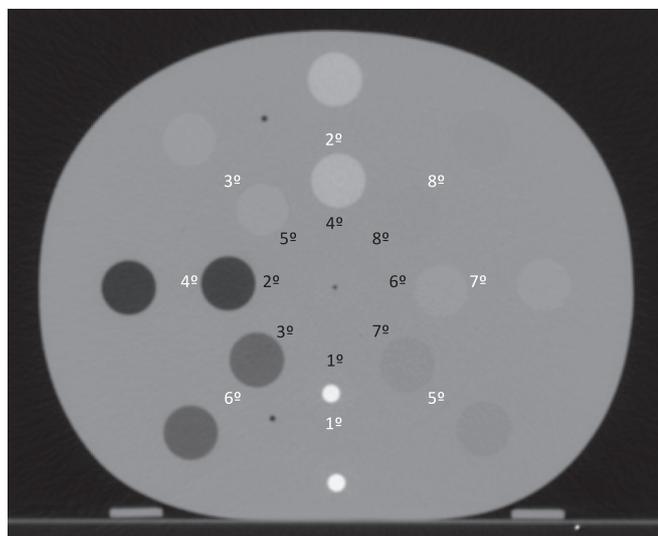


Fig. 4. Axial view of the CIRS phantom. In black the most repeated classification for contrast is shown. In white the most repeated classification for drop range is represented.

protocol. In addition, the edge characterization based on the values of contrast and drop range was comparable for all the protocols, according to the KRCC test. 22 out of the 37 radiomic features (59%) were comparable independently of the protocol.

3.2.2. Position dependency

For the same system and protocol (protocol C from PET/CT-NM at Center 1), different positions of the phantom inside the FoV were evaluated, positioning the phantom centered and off-centered. Neither the position of the phantom nor the ring (inner ring or outer ring in CIRS phantom) affected the comparability of the HU and the classification based on contrast values. However, position within the FoV showed a significant effect on drop range, reducing edge-sharpness, with an average 12% decay in the metric value when the phantom was placed off-centered, and the classifications obtained were not comparable based on KRCC, as seen in Table 2. 29 radiomic features (78%) were

Table 2

Comparability of calibration, edge characterization and radiomics measurements taken using protocol C from NM department at Center 1, used as reference protocol, and comparison between NM and RT department at Center 2. Bold used when the test showed the measurements where not comparable, left in white if they were.

Comparability to reference protocol (PET/CT-NM protocol C)		Calibration	Edge characterization		Radiomics
		HU (WRST)	Contrast (KRCC)	Drop range (KRCC)	Feature Extraction (max. 37, WRST)
Protocol dependency	Protocol	p-value			N° comparable features
PET/CT-NM	A	0.77	<0.001	0.03	24
	B	0.77	<0.001	0.01	30
	D	0.96	<0.001	0.02	24
Position Dependency	Compared Positions	p-value			N° comparable features
PET/CT-NM	Centered – In/Centered – Out	0.83	–	–	–
	Centered – In/Off-centered – In	0.92	0.002	0.18	29
	Centered – Out/Off-centered – In	0.83	–	–	–
	Centered – Out/Off-centered – Out	0.83	–	–	–
System dependency	Protocol	p-value			N° comparable features
CT-RT	A	0.80	0.002	0.98	26
	B	0.72	<0.001	0.55	30
	C	0.80	<0.001	0.11	32
	D	0.80	<0.001	0.90	28
CT-DR-1	QA	0.90	<0.001	0.40	18
	C'	0.96	0.013	0.11	32
CT-DR-2	QA	0.74	< 0.001	0.40	15
	C'	0.85	0.002	0.18	30
Center 2 Comparability	Protocol	p-value			N° comparable features
PET/CT-NM-F and CT-RT-F	C'	0.75	0.02	0.48	31

comparable independently of the position.

3.2.3. Acquisition system dependency

Measurements done with protocol C from NM department at Center 1 were compared with the ones realized with other CT systems at Center 1. For all systems and protocols, HU and edge contrast classification were comparable. Drop range classification was not comparable with other CT systems. Regarding radiomic features, different results were obtained for QA protocols compared to clinical protocols: 20 radiomic features (54%) were found to be comparable along all clinical protocols, compared to 13 radiomic features if QA protocols were implemented.

3.2.4. Comparability of calibration curves for CT systems in different institutions

To prove the feasibility of the proposed method for the comparison of CT performance across institutions, the calibration curve averaged over the four protocols at PET/CT-NM system of Center 1 (recommended calibration curve in Fig. 5) was compared to the calibration curves derived from the CT systems at NM and at RT department in the Center 2. Calibration curves for the two scanners in Center 2 were comparable, based on BA and WRST. Therefore, they were averaged to establish a recommended calibration curve for Center 2. This curve was comparable to the recommended curve at NM in Center 1. However, larger differences were obtained between the calibration curves of the different hospitals (NM (Center 1) and Center 2) with an average value of –7%, than between the calibration curves within the same institution (NM (Center 1) vs RT (Center 1)), with an average value of –4%, as shown in Table 3.

Dose difference due to the use of different calibration curves was assessed by calculating the gamma index for 4 different radiotherapy plannings in different cancer sites: brain, prostate, head and neck and lung. The results are shown in Table 4. As it is shown, the effect that the different calibration curves have in radiotherapy planning is negligible, Gamma (1 mm, 1%) > 99%. Nevertheless, differences increased (more points failed) for the calibration curve that showed larger relative differences in HU quantization (Table 3). A Dose-Volume-Histogram for Bronchial Carcinoma implementing the different calibration curves is shown in Fig. 6.

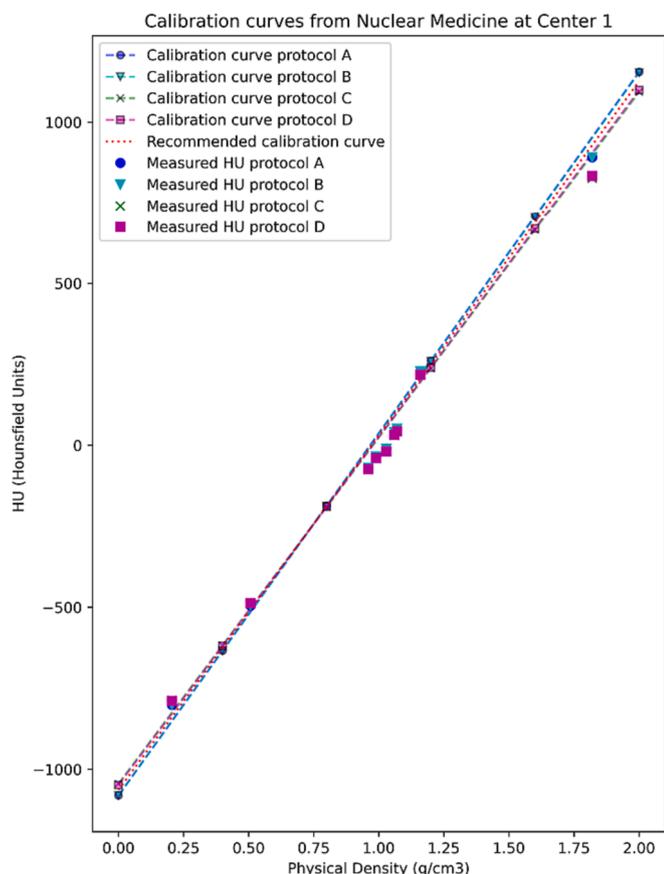


Fig. 5. Calibration curves for all measured protocols from NM at Center 1, recommended calibration curve and measured HU.

Table 3

Percentage deviation in electronic density derived from the recommended curve of the PET/CT-NM system at Center 1 with respect to the values derived from calibration curve for the RT system at Center 1 and the recommended calibration curve (averaged across NM and RT) in Center 2.

Deviation from NM (Center 1) recommended curve (%)		
Insert	RT (Center 1)	Center 2
Trabecular Bone	2	6
Breast	1	4
Muscle	1	5
Adipose	1	3
Dense Bone	3	10
Lung (Inhale)	-7	-14
Lung (Exhale)	-37	-89
Liver	1	5
Water	1	4
Mean deviation (%)	-4	-7

4. Discussion

This open-sourced automated QA software for CT images is fully developed in Python in conjunction with PyRadiomics for the radiomic features extraction and adapts to multiple phantoms. The software can assess the reproducibility of different CT metrics, including HU calibration, edge characterization and radiomic features.

The use of experimental phantoms is recommended by international committees [10,23,24] within QA programs for CT scanners. As to which phantom to use, there is not an agreement on the standard. Some of the proposed phantoms are the CatPhan Phantom [11,14,25], the American Association of Physicist in Medicine (AAPM) CT Performance Phantom [26–28] and the ACR CT Phantom [13,27,29,30]. Both ACR CT

Table 4

Gamma results comparing radiotherapy planning from the calibration curve from NM at Center 1 to the planning done with RT department curve and Center 2 curve. Results Gamma represents the percentage of voxels that were comparable and in parenthesis the number of voxels that failed. In Max Gamma the value in parenthesis represents the effect in dose that the different calibration curve has.

Accelerator	Results Gamma (1 mm, 1%)		Mean Gamma		Max Gamma	
	Center 1 (NM vs RT)	(Center 1 NM vs Center 2)	Center 1 (NM vs RT)	(Center 1 NM vs Center 2)	Center 1 (NM vs RT)	(Center 1 NM vs Center 2)
Brain (Clinac Static 6X 7Angle)	100% (13 failed)	100% (0 failed)	0.053	0.056	1.724 (0.005 Gy)	1.000 (0.005 Gy)
Prostate (Clinac VMAT 15X 2Arc)	100% (0 failed)	100% (20 failed)	0.071	0.258	1.000 (0.002 Gy)	1.261 (0.015 Gy)
Head and neck (Synergy Static 6X 3Angles)	100% (7 failed)	99.8% (7654 failed)	0.045	0.104	1.140 (0.006 Gy)	2.820 (0.016 Gy)
Lung (Clinac VMAT 6X 3Arc)	100% (0 failed)	99.3% (85147 failed)	0.075	0.341	0.900 (0.006 Gy)	2.036 (0.031 Gy)

Phantom and CatPhan Phantom are equipped with dedicated software for image analysis and have been implemented for QA in radiotherapy [29,31], but additional software is required for the radiomics features extraction. Furthermore, none of them has dedicated areas to simulate the electron density of the different tissues present in human anatomy. Other proposals [14–16] may feature an automatic QA process for a specific phantom, but do not adapt to other phantoms. The presented QA software is suitable for most phantoms, as it only requires reference segmentations and few adaptations on the software to work with a new phantom. Moreover, in contrast to other free access QA software, the rigid transformation implemented within our code allows the automation of the QA process. Therefore, once there are new images to be analyzed no additional work is needed other than adding them to the data base.

As a proof of concept, the proposed software has been tested with multiple images acquired from different CT systems and two different phantoms. No significant difference was observed between the different protocols regarding HU calibration, which is in line with what has been stated in different articles where neither the changes in the slice thickness [32] nor in the mAs [33] affect the measured HU. HU where comparable in different positions of the phantom, although moving the phantom away from the FoV lead to measurements slightly increasing [34]. In the comparison between Center 1 and Center 2, curves were comparable, although with a greater difference between Center 2 and NM department at Center 1 than between RT and NM at Center 1. It must be remarked that, because the phantoms employed were different depending on the institution, different density inserts were evaluated, and it could have contributed to the differences observed in the calibration curves. Moreover, based on the results of Gamma analysis, it could be concluded that calibration curve comparison by WRST and BA is a reasonable criterion for ensuring dose computation comparability. Regarding edges characterization metrics, both are pertinent as they have been used in the literature to study edges [35]. However, in our study drop range showed generally a poor comparability. Some studies emphasize the importance of image smoothing before edge detection as they are very noise sensitive [36]. This could have affected edge metrics reproducibility since no pre-processing was applied to the images.

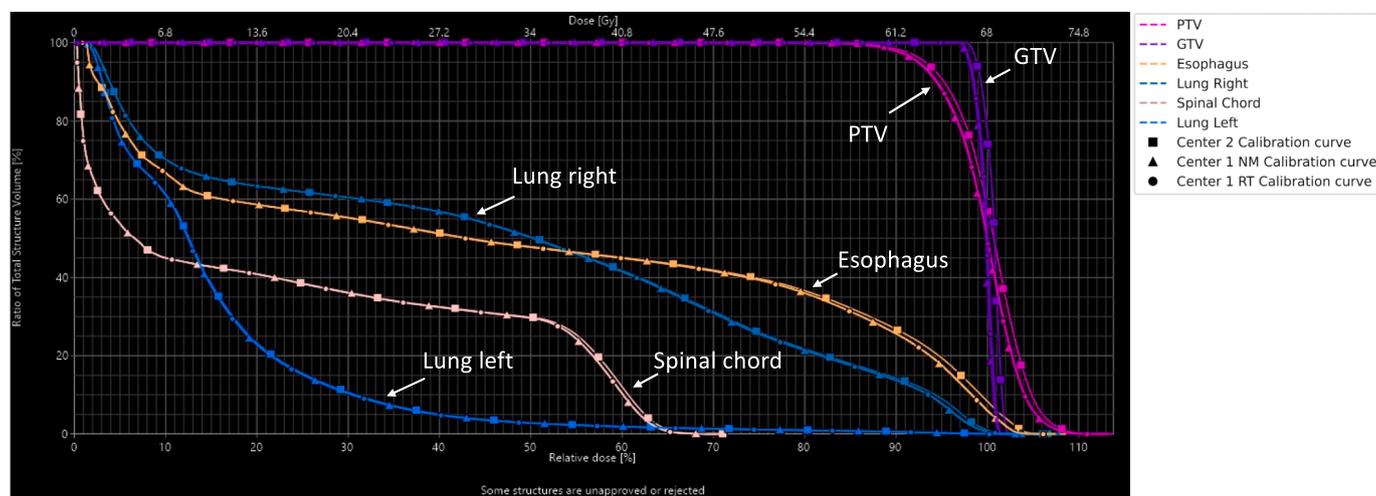


Fig. 6. Dose-Volume-Histogram for Bronchial Carcinoma for the same radiotherapy planning and CT image, but with different calibration curves (NM at Center 1, RT at Center 1, and Center 2).

Previous studies have evaluated radiomic features variability in PET [37–39] and MR [40,41] and it has been observed that voxel size showed a significant effect [37], which was also observed in this study.

Different QA routines for the evaluation of the CT system performance can be recommended. Firstly, a QA in the case of developing radiomic models with multicenter cohorts is recommended. By applying the software, it can be identified the different radiomic features that are comparable among all the CT systems. It will allow to implement in radiomic models only the most robust metrics, making possible not to mistake radiomic variation due to the equipment and not related with the patient, as in multiple studies CT scanner was found out to be a disruptive parameter in radiomics robustness [42]. In Table S2 at Supplementary Material a list of the radiomic features found out to be comparable among all CT scanners in this study is shown. Secondly, a consistency tests is proposed for QA of CT systems. In this case, 3 images should be taken for every clinical protocol. HU are expected to be comparable with respect to the last measurements based on WRST and calibration curves showing relative difference lower than 5%. Regarding edge characterization, contrast classification should be comparable to the last measurements based on KRCC test. If no differences in HU and edge contrast classification have been observed, those radiomic features previously determined to be robust are expected to remain so. We recommend performing the consistency test once per year for CT systems employed in diagnosis and every 6 months for CT systems involved in RT workflow, since in RT accurate HU quantification is needed to correctly compute the dose delivered to the patient and well-defined edges play an important role for an accurate and precise contouring of tumors and their surrounding organs-at-risk.

As a limitation of our study, no pre-processing was applied before metrics calculation and may be of interest to define the effect that image pre-processing could have in metrics reproducibility. Moreover, the CIRS phantom was not stored in facilities with controlled humidity and temperature, which could lead to a small absorption of water. However, while the phantom was not used to take measurements, it was kept in an insulated case proportioned by CIRS, so we expect the effect of humidity to be small. Furthermore, as the CIRS phantom has two different rings and the inner ring moves freely respect to the outer one, if it has not been fixed so that it does not move, a rigid transformation might not be able to successfully register the images and additional dedicated segmentations would be necessary.

5. Conclusions

An open-source software for the automatic evaluation of the analysis

and reproducibility of CT metrics has been developed. It has the capacity to adapt its functioning to multiple phantoms. The viability of the project has been tested with six different CT systems, two phantoms and two positions within the FoV, analyzing the metrics of the acquired images. Based on the obtained results assessing the reproducibility of the metrics different QA routines have been proposed.

Funding

Montserrat Carles was funded by the Conselleria de Sanitat Universal i Salut Pública from the Comunitat Valenciana. The funding sources had no involvement in the writing of the manuscript or in the decision to submit the article for publication.

Declaration of Competing Interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejmp.2023.103153>.

References

- [1] Smith-Bindman R, Kwan ML, Marlow EC, Theis MK, Bolch W, Cheng SY, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000–2016. *JAMA – J Am Med Assoc* 2019;322(9):843.
- [2] *Health at a Glance 2011*. OECD; 2011. doi:10.1787/health_glance-2011-en.
- [3] Fass L. Imaging and cancer: A review. *Mol Oncol* 2008;2(2):115–52. <https://doi.org/10.1016/j.molonc.2008.04.001>.
- [4] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;45(2):228–47.
- [5] Coolens C, Gwilliam MN, Alcaide-Leon P, de Freitas Faria IM, Ynoe de Moraes F. Transformational Role of Medical Imaging in (Radiation) Oncology. *Cancers (Basel)* 2021;13(11):25572557. <https://doi.org/10.3390/cancers13112557>.
- [6] Vijayakumar S, Yang J, Nittala MR, et al. Changing Role of PET/CT in Cancer Care with a Focus on Radiotherapy. *Cureus*. Published online December 22, 2022. doi: 10.7759/cureus.32840.
- [7] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;278(2):563–77. <https://doi.org/10.1148/radiol.2015151169>.
- [8] Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol* 2016;61(13):R150–66. <https://doi.org/10.1088/0031-9155/61/13/R150>.
- [9] Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to Radiomics. *J Nucl Med* 2020;61(4):488–95.

- [10] Mutic S, Palta JR, Butker EK, Das LJ, Huq MS, Loo L-N, et al. Quality assurance for computed-tomography simulators and the computed-tomography-simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 66. *Med Phys* 2003;30(10):2762–92.
- [11] Gulliksrud K, Stokke C, Trægde Martinsen AC. How to measure CT image quality: Variations in CT-numbers, uniformity and low contrast resolution for a CT quality assurance phantom. *Phys Med* 2014;30(4):521–6. <https://doi.org/10.1016/j.ejmp.2014.01.006>.
- [12] Sharp P, Barber DC, Brown DG, Burgess AE, Metz CE, Myers KJ, et al. 1. General Introduction. Reports of the International Commission on Radiation Units and Measurements 1996;os-28(1):1–5.
- [13] Dillon C, Breedren W, Clements J, et al. *Computed Tomography Quality Control Manual*; 2017.
- [14] Karius A, Bert C. QAMaster: A new software framework for phantom-based computed tomography quality assurance. *J Appl Clin Med Phys* 2022;23(4). <https://doi.org/10.1002/acm2.13588>.
- [15] Nowik P, Bujila R, Poludniowski G, Fransson A. Quality control of CT systems by automated monitoring of key performance indicators: a two-year study. *J Appl Clin Med Phys* 2015;16(4):254–65. <https://doi.org/10.1120/jacmp.v16i4.5469>.
- [16] Sun J, Barnes M, Dowling J, Menk F, Stanwell P, Greer PB. An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom. *Australas Phys Eng Sci Med* 2015;38(1):39–46. <https://doi.org/10.1007/s13246-014-0311-8>.
- [17] van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339.
- [18] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965;52(3–4):591–611. <https://doi.org/10.1093/biomet/52.3-4.591>.
- [19] Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. *Statistician* 1983;32(3):307. <https://doi.org/10.2307/2987937>.
- [20] Kendall MG. A new measure of rank correlation. *Biometrika* 1938;30(1–2):81–93. <https://doi.org/10.1093/biomet/30.1-2.81>.
- [21] Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Stat* 1947;18(1):50–60. <https://doi.org/10.1214/aoms/1177730491>.
- [22] Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys* 1998;25(5):656–61. <https://doi.org/10.1118/1.598248>.
- [23] Kuttner S, Bujila R, Kortesiemi M, Andersson H, Kull L, Østerås BH, et al. A proposed protocol for acceptance and constancy control of computed tomography systems: A Nordic Association for Clinical Physics (NACP) work group report. *Acta Radiol* 2013;54(2):188–98.
- [24] Samei E, Bakalyar D, Boedeker KL, et al. *Performance Evaluation of Computed Tomography Systems*; 2019.
- [25] Colli V, Mangini M, Strocchi S, Lumia D, Cani A, Boffano C, et al. Performance Assessment of Four 64-Slice Computed Tomographic Devices for a Typical Clinical Protocol. *J Comput Assist Tomogr* 2011;35(1):57–64.
- [26] Noh SS, Um HS, Kim HC. Development of Automated Quantitative Analysis Method in CT Images Evaluation using AAPM Phantom. *J Inst Electron Inform Eng* 2014;51(12):163–73. <https://doi.org/10.5573/ieie.2014.51.12.163>.
- [27] Lee KB, Nam KC, Jang JS, Kim HC. Feasibility of the Quantitative Assessment Method for CT Quality Control in Phantom Image Evaluation. *Appl Sci* 2021;11(8):3570. <https://doi.org/10.3390/app11083570>.
- [28] Park HJ, Jung SE, Lee YJ, Cho WI, Do KH, Kim SH, et al. Review of Failed CT Phantom Image Evaluations in 2005 and 2006 by the CT Accreditation Program of the Korean Institute for Accreditation of Medical Image. *Korean J Radiol* 2008;9(4):354.
- [29] Hobson MA, Soisson ET, Davis SD, Parker W. Using the ACR CT accreditation phantom for routine image quality assurance on both CT and CBCT imaging systems in a radiotherapy environment. *J Appl Clin Med Phys* 2014;15(4):226–39. <https://doi.org/10.1120/jacmp.v15i4.4835>.
- [30] Mansour Z, Mokhtar A, Sarhan A, Ahmed MT, El-Diasty T. Quality control of CT image using American College of Radiology (ACR) phantom. *Egypt J Radiol Nucl Med* 2016;47(4):1665–71. <https://doi.org/10.1016/j.ejrmm.2016.08.016>.
- [31] Davis AT, Palmer AL, Nisbet A. Can different Catphan phantoms be used in a multi-centre audit of radiotherapy CT image quality? *Phys Med* 2020;78:38–47. <https://doi.org/10.1016/j.ejmp.2020.09.003>.
- [32] Groell R, Rienmueller R, Schaffler GJ, Portugaller HR, Graif E, Willfurth P. CT number variations due to different image acquisition and reconstruction parameters: a thorax phantom study. *Comput Med Imaging Graph* 2000;24(2):53–8. [https://doi.org/10.1016/S0895-6111\(99\)00043-9](https://doi.org/10.1016/S0895-6111(99)00043-9).
- [33] Irsal M, Nurbaiti, Mukhtar AN, Jauhari A, Winarno G. Variation kVp and mAs on CT scan image quality using standard phantom. In: 2020:020039. doi:10.1063/5.0030320.
- [34] Zheng X, Gutsche L, Al-Hayek Y, Stanton J, Elshami W, Jensen K. Impacts of Phantom Off-Center Positioning on CT Numbers and Dose Index CTDIv: An Evaluation of Two CT Scanners from GE. *J Imaging* 2021;7(11):235. <https://doi.org/10.3390/jimaging7110235>.
- [35] Ahmad R, Ding Y, Simonetti OP. Edge sharpness assessment by parametric modeling: Application to magnetic resonance imaging. *Concepts Magn Resonance Part A* 2015;44(3):138–49. <https://doi.org/10.1002/cmr.a.21339>.
- [36] Noviana R, Febriani, Rasal I, Lubis EUC. Axial segmentation of lungs CT scan images using canny method and morphological operation. In: 2017:020022. doi:10.1063/1.4994425.
- [37] Carles M, Fechter T, Martí-Bonmatí L, Baltas D, Mix M. Experimental phantom evaluation to identify robust positron emission tomography (PET) radiomic features. *EJNMMI Phys* 2021;8(1):46. <https://doi.org/10.1186/s40658-021-00390-7>.
- [38] Carles M, Torres-Espallardo I, Alberich-Bayarri A, Olivás C, Bello P, Nestle U, et al. Evaluation of PET texture features with heterogeneous phantoms: complementarity and effect of motion and segmentation method. *Phys Med Biol* 2017;62(2):652–68.
- [39] Carles M, Bach T, Torres-Espallardo I, Baltas D, Nestle U, Martí-Bonmatí L. Significance of the impact of motion compensation on the variability of PET image features. *Phys Med Biol* 2018;63(6):065013. <https://doi.org/10.1088/1361-6560/aab180>.
- [40] Fernández Patón M, Cerdá Alberich L, Sangüesa Nebot C, Martínez de las Heras B, Veiga Canuto D, Cañete Nieto A, et al. MR Denoising Increases Radiomic Biomarker Precision and Reproducibility in Oncologic Imaging. *J Digit Imaging* 2021;34(5):1134–45.
- [41] Cattell R, Chen S, Huang C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. *Vis Comput Ind Biomed Art* 2019;2(1):19. <https://doi.org/10.1186/s42492-019-0025-6>.
- [42] Reiaz R, Abbas E, Famiyeh P, Rezaie A, Kwan JYY, Patel T, et al. The impact of the variation of imaging parameters on the robustness of Computed Tomography radiomic features: A review. *Comput Biol Med* 2021;133:104400. <https://doi.org/10.1016/j.combiomed.2021.104400>.