Atom Indonesia

Journal homepage: http://atomindonesia.brin.go.id

Automation of HU-RED Calibration: Development, Validation, and Monthly Consistency Testing in Radiotherapy

O. El-Mouden[1,2*], H. Sekkat[1,2], A. Bannan[1,2], R. El-Baydaoui[1], M. Bougteb[1], M. Mkimel[1], A. Khallouqi[1], M. Tahiri[1], Y. Berrada[2,3], F. Chaara[2,4], O. El-Rhazouani[1]

[1]*Laboratory of Sciences and Engineering of Biomedicals, Biophysics and Health, Higher Institute of Health Sciences, Hassan 1st University, Settat, Morocco*
[2]*International Hospital of Settat, Department of Radiotherapy, Settat, Morocco*
[3]*Faculty of Medicine and Pharmacy, Hassan II University of Casablanca, Casablanca, Morocco*
[4]*Faculty of Medicine, University of Seville, Seville, Spain*

ARTICLE INFO

Article history:
Received 31 October 2025
Received in revised form 7 April 2026
Accepted 9 April 2026

Keywords:

Radiotherapy
Quality assurance
CT calibration
Automation
Dose optimization

ABSTRACT

Radiotherapy delivers high doses of radiation to tumors, requiring careful optimization and quality management to maximize tumor control and minimize normal tissue damage. Quality Assurance (QA) protocols led by certified medical physicists are essential to maintain treatment accuracy within tight tolerances and prevent errors. In this study, we developed and validated an automated quality management procedure for the monthly calibration of Computed Tomography (CT) Hounsfield Unit to Relative Electron Density (HU-RED) conversion, as a case example of improving dose calculation accuracy in radiotherapy treatment planning through quality assurance optimization. A novel Python-based algorithm was implemented to extract CT calibration curves from phantom scans automatically, reducing human error and improving efficiency. The algorithm was tested on three different multi-material phantoms and compared its performance to manual analysis. The automated method significantly improved the accuracy of HU measurements and reproducibility of calibration, while reducing analysis time by over 90% compared to manual methods. All differences between automated and manual results were within 5 HU and statistically insignificant in terms of resulting electron density, with Pearson correlation R ≈ 0.999. Using the automated procedure, we established a monthly constancy test for CT calibration. Over 12 months of routine use, no significant drifts in the HU-RED relationship were detected, confirming the stability of the CT scanner's calibration within ±1-2% dose impact following the automated method. These results demonstrate how a systematic automated QA program ensures consistent dose calculation accuracy, aligning with international recommendations for radiotherapy safety.

© 2026 Atom Indonesia.
Published by BRIN Publishing. AI is ESCI and Scopus indexed. This is an open access article CC BY-NC-SA license (https://creativecommons.org/licenses/by-nc-sa/4.0/).

Article in Press

INTRODUCTION

In radiotherapy treatment planning, accurate conversion of CT Hounsfield Units (HU) to Relative Electron Density (RED) is essential for dose calculation in heterogeneous tissues [1-3]. A calibration curve mapping HU to RED is typically established during Treatment Planning System (TPS) commissioning by scanning a phantom with inserts of known density. Errors in this HU-RED correspondence can lead to significant dosimetry deviations. For example, HU shifts on the order of ±20 in soft tissues can alter the target dose by up to ~1%. Larger HU errors (±50) in lung or bone regions can yield local dose differences on the order of 1-2% [4,5], proving the need for precise calibration. It is generally recommended that CT scan parameters used for the phantom calibration (kVp, reconstruction filters, slice thickness, etc.) match those used clinically, since imaging protocol differences can influence the resulting HU values [5-8].

Initial HU-RED calibration is usually performed at TPS installation and verified periodically as part

*Corresponding author.
E-mail address: oussama.elmouden.doc@uhp.ac.ma
DOI: https://doi.org/10.55981/aij.2026.1841

of Quality Assurance (QA). However, many centers employ a single generic calibration curve (often at 120 kV) per CT scanner for all patients and anatomical sites, to minimize QA workload. This practice may be suboptimal if significantly different scan protocols (high-resolution bone kernels or different fields of view) are used, since applying a non-specific curve can introduce dose uncertainty. In particular, varying the reconstruction kernel without recalibrating the TPS can cause noticeable HU shifts that affect dose, one study showed that when HU changes stayed within ±20 for soft tissue, and ±50 for bone, the dose changes in targets remained under 1% [9], but using mismatched protocols can exceed these bounds. Similarly, phantom geometry and composition influence HU outcomes: using different calibration phantoms for the same scanner can yield HU differences of >20% for high-density materials. Goerlitz et al. reported that no single "universal" calibration curve applies to all scenarios, as body size and phantom material cause beam hardening effects that shift CT numbers. They estimated that using a head-sized against body-sized phantom (and kernels) could lead to ~1-2% dose calculation differences, and in a worst-case combined scenario, up to ~5% deviation [5]. These findings highlight the importance of performing protocol-specific HU-RED calibrations and regular consistency checks [10]. Related work in Atom Indonesia has also demonstrated the role of phantom-based CT performance checks, including CT-number linearity, and the use of CT-derived relative electron density in radiotherapy material characterization [11,12].

Automating the HU-RED calibration process is a promising strategy to facilitate frequent and protocol-specific QA without overburdening medical physicists. Recent work has explored automated QA procedures for HU-RED stability. For instance, Saborido-Moral et al. (2023) developed an open-source software to automatically evaluate HU-RED curve stability via repeated phantom measurements. Their study demonstrated excellent inter-protocol and inter-scanner reproducibility, finding no significant differences in calibration curves across varied acquisition conditions [13]. This suggests that modern CT scanners, when properly calibrated, maintain consistent HU-RED mappings across routine clinical protocols. In parallel, commercial solutions have emerged that integrate phantoms and analysis software for automating CT QA, for example, Sun Nuclear's RapidCHECK platform can automate CT-to-density calibration and trend analysis over time. Such platforms aim to standardize QA and provide intuitive feedback (immediate reports, tolerance checks, trending) for CT calibration tests [14]. Recent work from our group has also extended Python-based and QATrack+-integrated automation to CBCT HU-RED calibration, centralized EPID-based LINAC QA, and longitudinal LINAC/HDR quality-management dashboards [15-17].

Given these developments, the present work aims to design an in-house automated HU-RED calibration workflow and implement a monthly constancy test for the radiotherapy CT scanner. In this study, the development and validation of a Python-based algorithm for fully automated extraction of HU-RED calibration curves from phantom scans are presented, and its integration into a centralized QA database (QATrack+). The presented approach was evaluated by comparing it to manual analysis (considered the reference standard), and by testing its reproducibility across multiple phantoms and scanning protocols. The long-term stability of the calibration is also reported over 12 months of monthly checks. The goal is to demonstrate that automation can improve the accuracy, consistency, and efficiency of CT calibration QA while facilitating more frequent verification (monthly instead of yearly) with minimal additional workload. This aligns with QA guidelines such as AAPM TG-66 [18], which recommend regular monitoring of CT numbers for multiple reference materials (monthly) in addition to daily water checks. By automating the process and integrating it with QATrack+, we ensure reproducibility and continuous tracking of results, paving the way for proactive detection of any calibration drift and timely recalibration of the TPS if needed.

METHODOLOGY

CT scanner and imaging protocols

All scans were performed on the Discovery CT750 HD system CT simulator (GE Healthcare, Chicago, IL, USA) dedicated to radiotherapy planning. Six clinical scan protocols were considered: Head & Neck, Brain, Thorax, Abdomen, Pelvis, and Stereotactic Brain. These cover a range of patient imaging settings. To isolate the effect of anatomy-specific settings, we kept the tube voltage fixed at 120 kV for all protocols (as is standard for our planning scans) while varying other parameters (mA, rotation time, pitch, etc.) according to optimal clinical practice for each region. Table 1 summarizes the key acquisition parameters for four representative protocols (Pelvis used parameters similar to Abdomen and Stereotactic Brain similar to Brain). All scans were helical with a 512 × 512

ARTICLE IN PRESS

reconstruction matrix. Prior to scanning the phantoms, a blank scan (air scan) and a water phantom scan were performed for scanner calibration per manufacturer protocol. A uniformity check (scan without phantom) was also done for each protocol to verify consistent baseline HU and noise across the field. The phantom images were then acquired using the same parameters. After acquisition, all image series were exported in DICOM format for offline analysis.

Table 1. CT scanner acquisition parameters.

Parameter	Head & Neck	Brain	Thorax	Abdomen
Tube voltage (kVp)	120	120	120	120
Tube current (mA)	250 - 350 (auto)	220 - 300	150 - 400	200 - 400
Rotation time (s)	1.0	1.0	0.5	0.5
Helical pitch	0.984:1	0.531:1	0.984:1	0.984:1
Collimation (mm)	64 × 0.625	64 × 0.625	64 × 1.25	64 × 1.25
Slice thickness (mm)	0.625 - 1.25	0.625	1.25	1.25
Scan field of view (cm)	25 - 30	25	40 - 50	40 - 50
Display FOV (cm)	25	22	40	36
Reconstruction algorithm	Standard	Standard	Standard	Standard

Calibration phantoms

Three different heterogeneous phantoms were employed to cover a wide range of tissue densities and to validate the algorithm on various geometries: the CIRS Model 062M electron density phantom, a cylindrical water-equivalent phantom (CIRS Inc., Norfolk, VA) with interchangeable ring sections to simulate head or body sizes. It contains 10 cylindrical tissue-equivalent inserts of known electron density (RED). These represent a spectrum from lung to dense bone: air, lung inhale (~0.20 g/cc), lung exhale (~0.50 g/cc), adipose (~0.95 g/cc), breast tissue (~0.98 g/cc), water (1.00 g/cc), muscle (~1.04 g/cc), liver (~1.05 g/cc), trabecular bone (~1.17 g/cc, equivalent to 200 mg/cc Ca) and dense bone (~1.46 g/cc, ~1250 mg/cc Ca). The phantom's Plastic Water® housing (simulating soft tissue) allows for two setups (head or body diameter). This phantom is widely used for CT calibration in radiotherapy and provides a realistic HU-RED relationship due to its tissue-mimicking insert composition [19,20]. The second phantom, which was used, is the Easy Cube phantom, a modular cubic phantom (15 cm side) made of water-equivalent plastic (LAP GmbH, Germany). It accepts insert modules for various QA tests. We used the electron density module containing five cylindrical inserts: lung-equivalent (~0.20 g/cc), adipose (~0.97 g/cc), water (1.00 g/cc), muscle (~1.07 g/cc), and bone (~1.58 g/cc). These cover the typical electron density range encountered clinically. The compact design of Easy Cube yields higher intrinsic attenuation uniformity and is convenient for routine checks. Though it has fewer density points, it meets ESTRO and IEC recommendations and can be used for end-to-end tests from imaging to dose delivery [10]. The third one is Catphan 504 Phantom (CTP404 module), a well-known cylindrical QA phantom (The Phantom Laboratory, USA), commonly used in diagnostic and therapy CT QA. We used the sensitometry module CTP404, which has 7 inserts spanning from air to dense bone. The materials include air, polystyrene foam (lung-equivalent low density), low-density polyethylene (LDPE, lung-equivalent higher density), polystyrene (simulating fat/soft tissue), acrylic (simulating water or dense soft tissue), Delrin® (acetal, simulating spongy bone), and Teflon® (PTFE, simulating cortical bone). Each insert is ~3 cm in diameter, arranged in a circular pattern [21]. This phantom provides a more limited but industry-standard HU linearity check. We included it to test the algorithm's robustness in identifying fewer/smaller inserts with closer spacing.

The known physical density and electron density values for the CIRS and Catphan inserts were taken from manufacturer specifications. Briefly, the CIRS inserts span, with electron densities from 0 (air) to ~4.86 × 10^{23} e⁄cm³ (dense bone), corresponding to relative electron densities 0.0 to ~1.456. The Catphan CTP404 inserts span ~0 to 6.24 × 1023 e⁄cm³ (relative RED 0.0 to ~1.867). These values were used as a reference RED for plotting calibration curves. Each phantom was scanned with the same CT protocol (120 kV, standard kernel) to allow inter-comparison of the resulting HU-RED curves.

Automated HU-RED calibration algorithm

A dedicated Python script was developed to automatically process the phantom CT images and extract the HU-RED calibration data. The algorithm leverages the open-source Pylinac library (v3.0+, by James M. Kerns), which provides tools for analyzing CT QA phantoms [22]. Pylinac includes an automated Catphan analysis module. We extended and customized this module to also recognize and analyze the CIRS 062M and Easy Cube phantoms. The general workflow of the algorithm is illustrated in Fig. 1. The steps are as follows: firstly, DICOM images are imported, and the axial CT image series of the phantom is

Article In Press

loaded. Basic preprocessing is done, including sorting slices and checking image uniformity (verifying no truncation or abnormal artifacts). Secondly, the code determines which phantom was scanned by analyzing the image geometry and the insert pattern. For example, the CIRS phantom has a larger diameter and 10 inserts spanning the full circle, whereas Catphan's CTP404 module has a distinctive central outline with 7 inserts, and Easy Cube has a small square outline. We also detect orientation using known markers (air insert position or phantom engraving) to establish a reference angle. Given the identified phantom type, the algorithm applies a specific strategy to find each insert. For CIRS, we detect the phantom outer contour and then locate inserts at expected radial distances and angles (accounting for the known configuration of the 10 insert holes). For Catphan, we use Pylinac's approach of finding the high-contrast registration holes and the pattern of inserts. For EasyCube, with a linear arrangement of inserts, a simple template matching or intensity profile analysis along the known insert plane is used. Image filtering (thresholding, edge detection) is applied as needed to enhance insert visibility. Once each insert ROI is identified, the algorithm computes the mean HU value within that ROI. As shown in Fig. 2, the automated algorithm identifies the inserts and defines the corresponding regions of interest (ROIs) for HU extraction across the different phantom geometries. The ROI size is predefined (a circular region ~50-70% of insert diameter) to avoid edge artifacts from the high-contrast boundaries [5]. The standard deviation of HU in the ROI is also recorded as an uncertainty estimate. For each material insert, we now have an average HU_measured. The measured HU values are paired with the known electron densities (or relative electron densities) of the corresponding insert materials. The conversion curve is then generated. Typically, a piecewise linear interpolation between data points is used, or a polynomial fit if desired. In our case, we store the raw point pairs (HU, RED) for flexibility. This constitutes the calibration curve for that scan. The newly obtained curve is automatically compared to a baseline or reference calibration curve for the scanner. The reference can be the initial commissioning calibration or the previous month's result. Key indicators are calculated, such as the difference in HU for each material (ΔHU), the difference in RED for each material (ΔRED), and the correlation coefficient between the new and reference curves. These are used to assess if the new measurement is within tolerance. We set tolerance thresholds per material type (derived from known acceptable HU variability): ±30 HU for water-equivalent, ±50 HU for lung and dense bone, ±40 HU for adipose/soft tissue ranges [4]. We also require the overall curve correlation to remain >0.99. These tolerances were defined based on literature reports of dose sensitivity [4] and our own commissioning experience. The results, including the measured HU values for each insert, the generated calibration curve (and fit parameters if any), and the QA comparison metrics, are saved. The automation is integrated with QATrack+, an open-source QA database platform widely used in radiotherapy clinics [23-25]. Each time the script is run, it uploads the results to QATrack+ via its API. The data are stored with the date, scanner protocol, phantom used, and all measured values. In QATrack+, custom composite tests compute if tolerances are exceeded and trigger alerts accordingly. This provides a permanent, centralized record of the CT calibration history.

The entire process from image import to result output is fully automated, requiring no user intervention beyond initiating the script once the phantom scan is available. The algorithm runtime was measured on a standard PC (Intel Core i7 CPU): processing one scan takes on the order of 1 minute or less, an important improvement over manual analysis, which can take ~15-20 minutes per phantom. The Python code was modular, allowing easy adaptation to additional phantoms in the future by adding their geometric template.

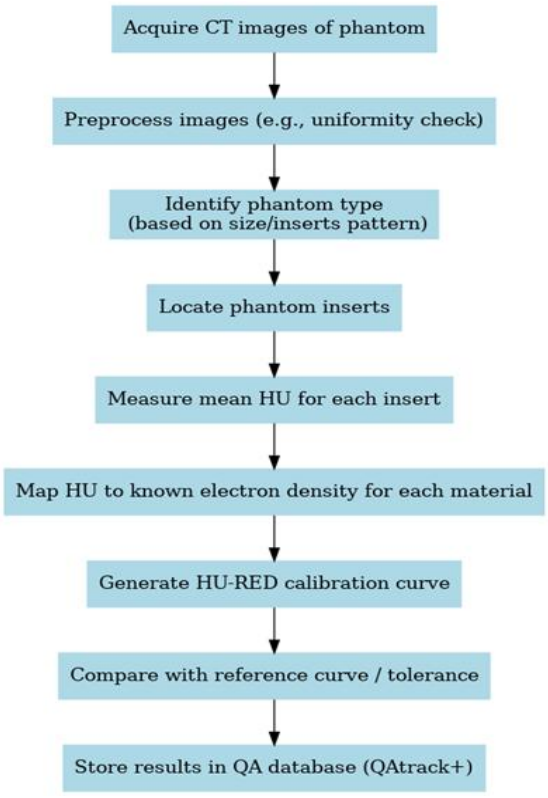
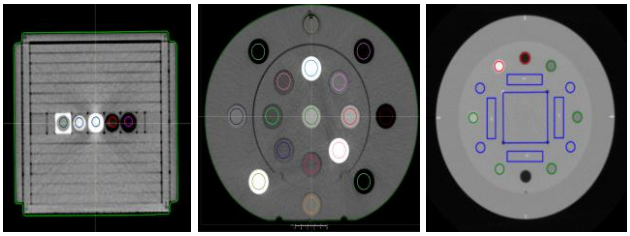


Fig. 1. Automated algorithm flowchart for HU-RED calibration.

Article in Press



a b c

Fig. 2. Automated detection and segmentation of inserts for different phantom types: (a) Easy Cube phantom, (b) CIRS 062M phantom, and (c) Catphan 504 phantom. Detected inserts and corresponding regions of interest (ROIs) used for HU extraction are overlaid on representative axial CT images.

Manual analysis for validation

To validate the automated method, we performed a traditional manual analysis for each phantom scan to serve as the ground truth reference. A medical physicist analyzed the DICOM images either on a treatment planning workstation or using ImageJ (NIH) image analysis software, following the typical procedure recommended by the IAEA (IAEA-TECDOC-1583 report) for HU-RED calibration [26]. This involved placing circular ROI markers (~1-2 cm diameter) in the center of each phantom insert on a representative axial slice and recording the mean HU. The measured HU values were then manually paired with the known RED of each insert to construct a calibration curve. This manual process was repeated for each phantom and each imaging protocol as needed. To ensure consistency of monthly measurements, all analyses were performed under standardized conditions. The same CT scanner, phantom configuration, acquisition protocol, and reconstruction settings were used throughout the study period. Phantom positioning was carefully controlled, and no additional preprocessing or filtering was applied prior to analysis. Routine scanner quality assurance procedures were maintained to minimize external sources of variability. While tedious and subject to inter-user variation, this method is the current standard in many clinics and was used as the benchmark for our algorithm's accuracy. Additionally, to assess short-term reproducibility, each phantom was scanned multiple times in one session. For a given protocol, we performed repeated scans (with the phantom repositioned between scans) and analyzed each with both the automated and manual methods. This tested the consistency of HU measurements and the robustness of insert detection. We also conducted a deliberate variation test: the CIRS phantom was scanned under slightly altered conditions (using a different reconstruction filter or off-centering the phantom by a few cm) to simulate potential errors or changes. Both methods analyzed these to see if they detect the induced changes in the calibration curve.

To ensure consistency between the manual and automated measurements, both methods were applied to the same original DICOM image datasets acquired under identical scanning conditions. No additional preprocessing, filtering, or image modification was applied prior to either analysis. The automated algorithm processed the original images directly, whereas manual measurements were performed on the same datasets using standard clinical software. ROI placement strategies were also kept as comparable as possible: manual ROIs were positioned centrally within each insert, while automated ROIs were defined proportionally to the insert size (typically 50 - 70% of the insert diameter) to sample the same central region while minimizing edge effects and partial-volume artifacts. This ensured that any observed differences between the two methods were attributable primarily to the analysis workflow rather than to differences in image preprocessing or ROI definition.

Integration into QAtrack+ platform

A major component of this work was integrating the automated calibration into our departmental QAtrack+ system. QATrack+ is an open-source web platform for managing QA data, allowing tests to be stored, trended, and reviewed easily [23-25]. We configured QATrack+ with a custom test list for "CT Simulator Monthly HU-RED Calibration." The automated script communicates with QATrack+ via its REST API to populate this test each time the calibration phantom is scanned. The following data are recorded for each monthly test: the scanner protocol used, each insert's measured HU (by the algorithm), the computed curve parameters, and key comparison metrics (max ΔHU, ΔRED, correlation to reference, etc.). QATrack+ was set to automatically apply the tolerance criteria described above: if any insert's HU deviation exceeds its threshold (e.g., water >±30 HU) or if the curve correlation drops below 0.99, the test is flagged out of tolerance. We implemented automated email notifications and dashboard alerts in QATrack+ for any out-of-tolerance result. The platform's trending feature was utilized to plot each material's HU over time, enabling visual tracking of any gradual drift. For example, we monitor the monthly HU of the dense bone insert to catch any slow decline that might eventually approach the ±50 HU tolerance. By centralizing the results in QATrack+, we ensure complete traceability of the CT calibration over time, elimination of manual transcription errors, and easy access for all physicists in the team. This integration

Article in Press

effectively closes the loop: once the phantom is scanned, all analysis and logging to QA records happen automatically, making a monthly calibration program feasible without significant staffing burden. While our current workflow stops at QA monitoring, we envisage a future link from QATrack+ to the TPS: if a persistent calibration shift is observed, the system could assist in generating an updated HU-RED table and notifying the physicist to import it into the TPS (ensuring the planning system stays in sync with the scanner). Such end-to-end integration, from scanner to TPS, would further enhance safety by minimizing the latency and human error in maintaining an accurate CT calibration.

RESULTS AND DISCUSSION

The automated algorithm's HU measurements showed very high agreement with manual ROI measurements, as shown in Fig. 3. For all phantoms and materials, the differences between automated and manual HU values were on the order of a few HU. The mean absolute error per insert (averaged across materials) was ~3-5 HU for the automated method vs ~6-8 HU for manual, relative to the expected reference values. The dashed line is the identity line (perfect agreement). The data points lie very close to the identity line, indicating that the automated analysis reproduces the manual results within measurement uncertainty. The largest outlier observed was on the order of ~10 HU difference for a high-density bone insert in one case, which is still within 1% of the material's HU value (~1000 HU). Statistical analysis confirmed no significant bias between methods (paired t-test $p \gg 0.05$; differences were random around zero). The results in Table 2 validate that the algorithm can replace manual ROI placement without loss of accuracy, even for challenging low-contrast inserts. We also compared the resulting calibration curves produced by each method. Each automated HU-RED curve was essentially indistinguishable from its manual counterpart. The maximum deviation in RED at a given HU point was under 0.005 (relative units). This negligible difference implies the dose calculation would be unaffected. We evaluated the impact by comparing dose calculations using automated and manual calibration curves in a TPS for a sample treatment plan: dose to a target in a heterogeneous phantom differed by <0.2% between the two, which is within typical Monte Carlo statistical uncertainty. Therefore, the automated workflow produces calibration curves of equivalent quality to the standard manual method. The automated method demonstrated improved precision in repeated measurements. With back-to-back scans of the same phantom, the standard deviation of measured HU for each insert was very low (on the order of 1-2 HU), corresponding to a coefficient of variation <1%. Manual analysis, in contrast, showed 2-3× higher variation, likely due to slight differences in ROI placement between repetitions. For example, over 5 repeat scans of the CIRS phantom, the automated HU for water was 0±2 HU (CV (coefficient of variation)=0.5%) while manual ROI readings ranged 0±5 HU (CV=1.5%). This improvement in consistency is partly because the algorithm defines ROIs identically each time, whereas human placement can shift. The table summarizes key comparison metrics across methods and phantoms: The automated analysis slightly outperforms manual in terms of HU accuracy (smaller errors and max deviations) and greatly outperforms in consistency and speed. ROI detection was nearly 100% reliable; the algorithm correctly identified all inserts in all scans, whereas manual analysis occasionally missed a very faint low-density insert on the first try (hence slightly lower "success" percentages, though ultimately the user can find it).

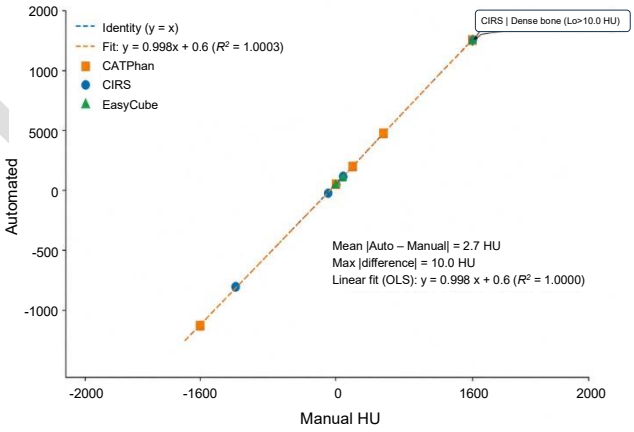


Fig. 3. Correlation of automated and manual HU measurements, including all phantom inserts (Catphan, Cirs, and Easy Cube).

Table 2. Automated against manual analysis performance (averaged across all inserts).

Metric	Phantom	Automated (Python)	Manual (ImageJ)	Relative improvement
HU Mean Error (HU)	Catphan	±3.2	±5.7	+44% accuracy (p<0.01)
	CIRS 062M	±4.1	±6.3	+35% (p<0.01)
	Easy Cube	±3.8	±5.9	+36% (p<0.01)
HU Max Deviation (HU)	Catphan	8.5	14.2	40% lower (p<0.01)
	CIRS 062M	9.3	15.8	41% lower (p<0.01)
	Easy Cube	8.9	16.3	45% lower (p<0.01)
Coefficient of Variation (%) *(Reproducibility over n scans)*	Catphan	0.8	2.1	62% lower (p<0.001)
	CIRS 062M	1.2	2.6	54% lower (p<0.001)
	Easy Cube	1.1	2.3	52% lower (p<0.001)
ROI Detection Success (%)	Catphan	99.7	97.5	– (p=0.03)
	CIRS 062M	98.9	96.8	– (p=0.04)
	Easy Cube	99.5	98.2	– (p=0.08)
Analysis Time (minutes)	Catphan	0.8	15.3	~95% faster (p<0.0001)
	CIRS 062M	1.2	22.6	~95% faster (p<0.0001)
	Easy Cube	0.9	17.2	~95% faster (p<0.0001)

Article Press

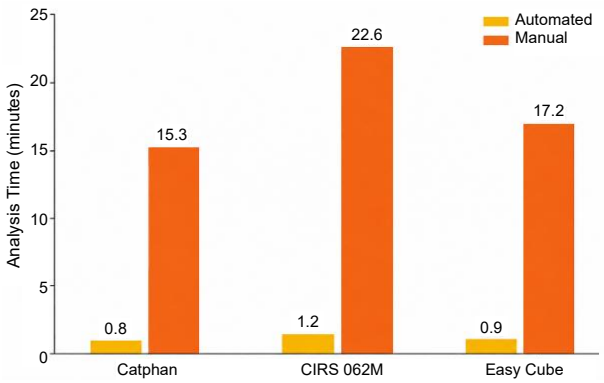


Fig. 4. Automated against manual analysis time.

Perhaps the most important improvement is in analysis time. The automated script processed a phantom scan in under 1-2 minutes, compared to 15-20 minutes for a careful manual analysis (including data recording). Figure 4 highlights this difference for the three phantoms. Automated processing is ~15 - 20× faster than manual ROI measurement for each phantom. This ~95% reduction in analysis time significantly decreases the workload for medical physicists and enables high-frequency QA. What used to be a laborious task now runs in the background with negligible time cost. In summary, the validation confirms that the automated workflow can replace manual measurements, yielding equivalent HU-RED calibration curves with improved reproducibility. All further analyses (inter-phantom, longitudinal stability) were therefore conducted using the automated method.

Inter-phantom comparison of calibration curves

Using the automated tool, we extracted the HU-RED calibration curves from each of the three phantoms under identical scan settings (120 kV, standard recon). Figure 5 shows the resulting curves (HU vs. relative electron density) for the CIRS 062M, Easy Cube, and Catphan phantoms. Each point corresponds to an insert material. Note the divergence at high densities among the phantoms (a): Low-density zoom "Zoom - low densities (RED ≈ 0.00-1.10): air → water" (b): High-density zoom"Zoom - high densities (RED ≈ 0.95-1.65): bone / dense plastics".

All three curves are reasonably consistent in the low- to mid-density range: from lung-equivalent up to water and soft tissue (~1.0 RED), the data overlap closely. For instance, the water-equivalent inserts in all phantoms registered ~0 HU, and muscle/soft-tissue equivalents (~1.04-1.07 RED) were in the 20-40 HU range for all, showing negligible differences. This indicates that for soft tissues, the phantom material compositions (whether epoxy, polyethylene, etc.) mimic water sufficiently well that the scanner yields similar HU.

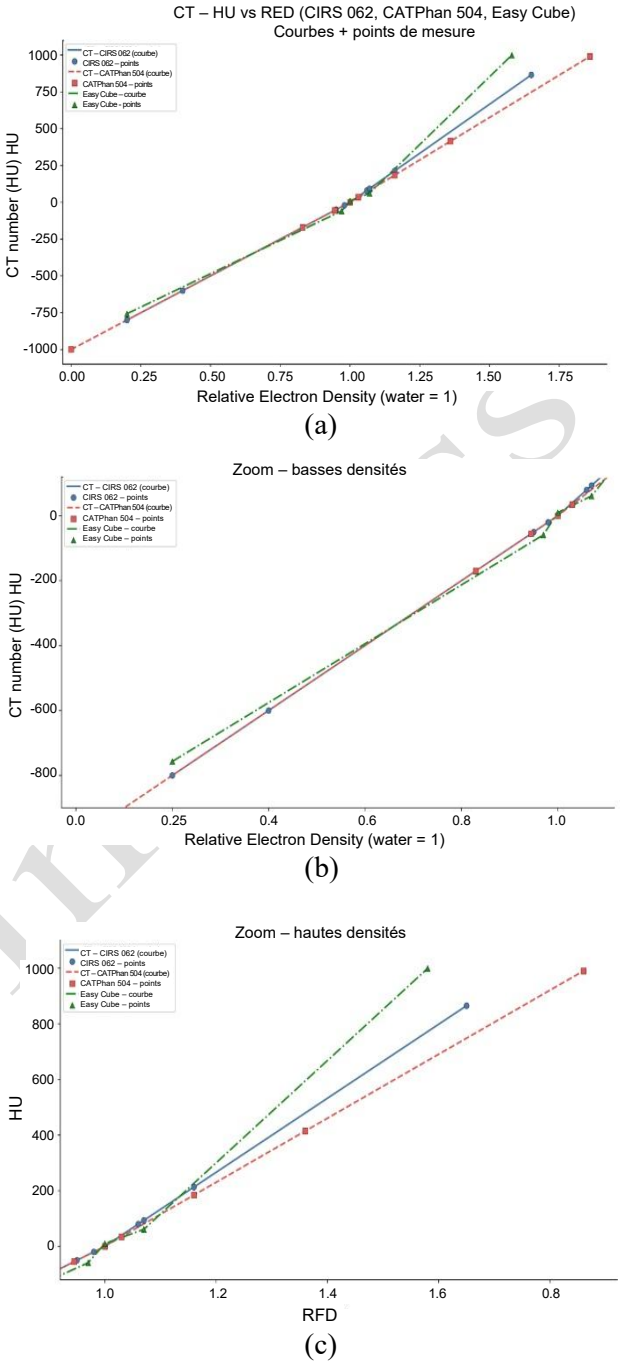


Fig. 5. CT calibration curves relating Hounsfield units to relative electron density. (a) Complete HU-RED calibration curves obtained using the CIRS 062, Catphan 504, and Easy Cube phantoms, with corresponding measurement points. (b) Zoomed view of the low-density region, highlighting the agreement between phantom materials in the lung- and soft-tissue-equivalent range. (c) Zoomed view of the high-density region, showing differences between calibration curves at higher relative electron densities. Overall, the curves show consistent HU-RED behavior across the clinically relevant density range, with larger deviations observed at high densities.

However, at the extremes of the density range, we observe phantom-dependent variations. The largest discrepancies are in the high-density (bone) region. The Catphan's cortical bone substitute (Teflon, RED≈1.87) yielded a much higher HU (~950 HU) than the CIRS's dense bone

insert (epoxy with RED≈1.46, ~740 HU). The Easy Cube's bone insert (RED≈1.58) gave an intermediate HU (~800 HU). Consequently, for materials representing cortical bone/dentin, the measured HU differed by on the order of 200 HU between phantoms. In relative terms, the Catphan vs CIRS high-density point differed by about 22-25% in HU (950 vs 740), consistent with prior reports [9]. These differences arise from the phantoms' material composition and size: Teflon has a higher atomic number effective (Z_{eff}) and causes more beam hardening, pushing CT numbers higher. The CIRS bone insert, while physically dense, has additives to match electron density with human bone, resulting in a lower Z_{eff} and thus lower HU. Additionally, the larger diameter of the CIRS phantom (body configuration) introduces more beam hardening, mainly, which can reduce HU for internal inserts compared to the smaller Catphan (this effect is less pronounced for bone against soft tissue) [5]. The Easy Cube's intermediate size and composition yield bone HU in between.

In the low-density range, a slight phantom effect is also noticeable for lung-equivalent foam. The CIRS lung inhale insert (RED≈0.20) measured around –800 HU, whereas Catphan's lowest density foam (LDPE or polystyrene foam, RED≈0.30–0.35) measured –700 to –600 HU. Easy Cube's lung (~0.20) was ~–750 HU. Some of this variation is due to the exact density differences (and possibly how well the scanner's beam hardening correction is calibrated for different phantom sizes). Nonetheless, all are within a few hundred HU of air (–1000 HU). The mid-density points (fat ~0.95, water 1.0) were virtually identical across phantoms (all gave HU ≈ –100 for fat and ~0 for water), which is reassuring. These findings reinforce that while all phantoms are suitable for HU-RED calibration, there will be systematic offsets, particularly at high HU, if one uses a different phantom than the one with which the TPS was commissioned. Jaafar et al. noted the same issue: a ≥22% HU difference for high-density inserts between Catphan and CIRS led to significant differences in the RED calibration if not accounted for [9]. In practice, if a clinic switches phantoms, a new calibration curve should be obtained rather than assuming the previous curve applies. Our data also suggest that a calibration curve derived from one phantom (e.g., Catphan) might slightly overestimate bone density if applied to patient scans, compared to a curve from a tissue-mimicking phantom (CIRS). However, the dosimetry impact of these discrepancies is modest. Using two quite different curves (Catphan vs CIRS) in a TPS dose calculation led to a <2% difference in dose for a bone-embedded target (in agreement with Goerlitz's ~1–2% estimate for phantom differences [5]). Still, consistency is key: it is recommended to use the same phantom model for periodic checks as was used in commissioning, or at least be aware of the offset and adjust tolerance accordingly if using a different phantom. In summary, the inter-phantom comparison showed excellent agreement in HU-RED calibration for low and medium densities, and some expected divergence at high densities due to material differences. No significant differences were observed in the shape of the calibration curve between phantoms aside from the vertical offsets of those extreme points. This indicates our automated algorithm handled all phantom types correctly and that the CT scanner's response is stable; the differences seen are attributable to phantom material properties. All curves remained linear and monotonically increasing, as expected, and no phantom indicated an out-of-trend result for the scanner.

Influence of scan protocol (Inter-protocol consistency)

Given that all our protocols used 120 kV, we anticipated minimal differences in HU-RED calibration between the Head & Neck, Thorax, etc. protocols. Indeed, our measurements confirmed that the scan protocol did not significantly affect the calibration curve, so long as the kV and reconstruction kernel were the same. The Head & Neck vs. Abdomen protocols, for example, produced virtually identical HU values (within ~5 HU) on the CIRS phantom, despite differences in FOV and mA. This agrees with the findings of Saborido-Moral et al. that HU calibration is comparable across typical clinical protocols [13]. One exception we tested was the Brain vs. Body kernel: using a sharp high-resolution kernel (Brain) versus a standard soft kernel (Body) did introduce small differences; high-contrast inserts like bone showed ~20 HU higher with the sharp kernel, consistent with known kernel-induced HU [4]. However, in radiotherapy planning, it is standard to use the same "soft tissue" reconstruction for all anatomical sites to preserve HU accuracy. Our institution follows this practice,

Article In Press

so routine head vs body scans share the same kernel. Therefore, within our typical protocols, we found the HU-RED curve to be protocol-independent. This justifies using a single reference curve per scanner (at 120 kV) as long as all scans use the calibration-specific kernel and parameters. It also validates that our monthly checks (which we performed using the Abdomen protocol) are applicable to the Head & Neck protocol, etc. If in the future a different kV or kernel is introduced clinically, we would simply run the automation on that protocol and generate a separate curve (the script can easily accommodate multiple reference curves for different techniques).

Monthly calibration stability and QA tracking

After deploying the automated calibration and QATrack+ integration, Monthly HU-RED cons istency tests were conducted over a 12-month period. Each month, the CIRS 062M phantom was scanned using the standard abdomen protocol, and the automated analysis recorded the results to QATrack+. All monthly data fell within our pre-set tolerances, and no significant drift was observed in the calibration curve. The observed stability should be interpreted in the context of controlled acquisition and analysis conditions maintained throughout the study period. Figure 6 shows the trend of HU for selected insert materials across the year. Red dashed lines indicate tolerance limits for each (water ±30 HU, lung ±50 HU, bone ±50 HU). All measurements remained within tolerance. The water-equivalent insert remained very close to 0 HU each month (within ±10 HU). The lung insert varied within roughly ±10-15 HU around its baseline (~–800 HU), and the dense bone insert within ±15-20 HU around its ~740 HU baseline. These fluctuations are minor relative to the tolerances (red lines) and show no obvious upward or downward trend. For instance, the dense bone HU was 750 HU in month 1, dipped to ~730-735 HU in mid-year, and was ~745 HU by month 12, all within a 3% band. The correlation coefficient of each monthly curve to the baseline curve stayed >0.999, well above the 0.99 threshold. This indicates the CT scanner's calibration was stable over the year.

Importantly, no QA alerts were triggered by QATrack+ during the 12 months. The system automatically marked each test as "Within Tolerance" (green status), and no email notifications of deviation were sent. This gives confidence that the scanner did not experience any significant calibration shift (such as might occur from X-ray tube aging or a major hardware change). It should be noted that routine maintenance (including tube changes) did not occur during this interval; had it occurred, we might have expected a one-time jump in HU, which this program would catch. Our proactive monitoring aligns with published QA recommendations that suggest verifying CT numbers on multiple materials monthly, in addition to daily water HU checks. AAPM TG-66 [18], for example, suggests a reduced set of inserts monthly and a full phantom scan annually. With our automation, a full multi-insert check monthly became practical, and it indeed confirmed consistent performance over time. The integration with QATrack+ proved valuable in visualizing its stability. We could easily generate trend plots for each insert's HU via the QATrack+ web interface, and the data are stored for future reference or audits. The trending helps to identify any subtle drift, for example, if the lung HU had shown a steady downward trend of ~5 HU per month (even if still within tolerance each month), we would notice that on the plotted graph and could investigate proactively. In our case, the trends were essentially flat. The QA system provides an extra layer of safety; had any monthly data point exceeded tolerance (say water HU = +35 HU one month), QATrack+ would have immediately flagged it, and we could schedule an investigation or re-calibration before any potential dosimetry impact.

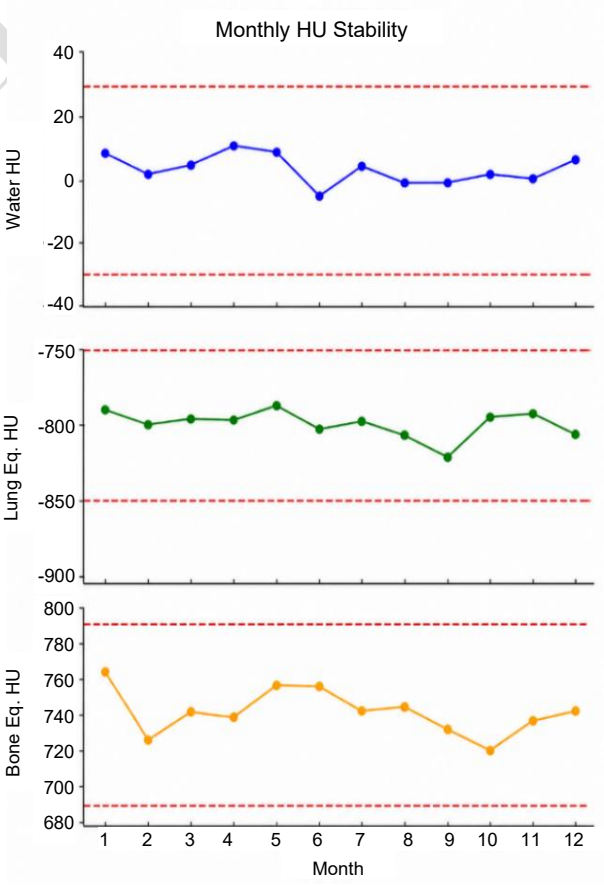


Fig. 6. Monthly stability of CT numbers for key reference materials: (top) water-equivalent, (middle) lung-equivalent, (bottom) dense bone-equivalent.

Article in Press

At the end of 12 months, we performed an annual verification using all three phantoms and compared them to the initial commissioning data. The results showed no statistically significant change in the HU-RED calibration curve from the initial baseline (all $p>0.5$), confirming the scanner's long-term constancy. This exercise demonstrates how an automated monthly program can ensure there are no surprises at annual QA; any issue would have been caught earlier. Our experience also highlights that the tolerance values chosen (±30 HU for water, ±50 HU for lung/bone) were appropriate, as the normal variation observed was well within these ranges. These tolerances are in line with literature, for example, Davis et al. note that HU changes of ±20 in soft tissue and ±50 in bone/air yield <1% dose differences [4], which is generally clinically insignificant. Therefore, our threshold criteria would only trigger if a change likely to cause >1% dose effect occurred, which never happened. Keeping the criteria somewhat broad avoids false alarms - small day-to-day HU fluctuations (5-10 HU) are expected and not dosimetrically critical.

This study demonstrates a successful implementation of automated HU-RED calibration in a radiotherapy context, yielding improvements in efficiency and consistency that can directly benefit QA practice. The automated algorithm was shown to be as accurate as manual analysis while virtually eliminating user variability and reducing analysis time by an order of magnitude. By integrating the tool with a QA database (QATrack+), we achieved a robust, reproducible workflow for regular calibration checks that was previously impractical to conduct with high frequency.

The ability to easily generate and track calibration curves has several implications. First, medical physicists can now verify their CT scanner's HU-RED calibration monthly (or after any major event) with minimal effort, rather than perhaps once a year due to time constraints. This greatly increases the likelihood of catching any calibration drift early. Although our observed calibration was very stable, not all systems may be; an aging X-ray tube, a change in scanner beam filtration, or a firmware update could alter HU subtly. With automated monthly checks, one can ensure the TPS conversion table remains valid or update it promptly if needed, therefore maintaining dosimetric accuracy for patient plans [5]. Although the automated workflow reduced operator-related variability, the consistency of monthly results may also have been influenced by external factors such as scanner, phantom positioning, acquisition settings, and reconstruction conditions. In this study, these factors were controlled as much as possible by maintaining the same CT scanner, phantom configuration, acquisition protocol, and reconstruction parameters throughout the study period, together with routine scanner QA procedures. Our approach effectively introduces a new QA test to the monthly schedule, but the automated approach introduces minimal additional workload. This is a clear benefit for patient safety, as even small dose discrepancies in critical cases (lung SBRT near spinal cord) could be a concern if HU-RED were off by a large amount.

The time saved is significant. In our case, analyzing three phantoms manually would take ~1 hour of a physicist's time; the automated script handles all three in ~3 minutes. Over a year, if done monthly, that's 12 hours vs 36 minutes, a substantial resource difference. This efficiency might encourage sites to expand their calibration to multiple protocols or phantoms. For example, one could check 120 kV and 140 kV calibrations, or both head and body phantoms, whereas before one might do just a single phantom at a single energy due to effort. The open-source nature of our solution means no additional cost for software, making it accessible to institutions that might not afford commercial tools.

Our findings corroborate those of Saborido-Moral et al., who also found that automation yields high reproducibility and that HU-RED curves were consistent across protocols and scanners [13]. They further extended automated QA to texture analysis, in our work, we focused on the HU-RED aspect specifically, but similarly leveraged open-source code. Other groups have also reported on automated CT QA. For instance, an effort by Taylor et al. introduced QATrack+ as a framework for routine imaging QA, allowing custom scripts for tests. Our integration is an example of harnessing such a framework for a specific advanced test. Regarding the phantom differences we observed, Jaafar et al. discussed the influence of kV and phantom on HU-RED curves and recommended using tissue-equivalent materials for calibration [9]. Our results agree that tissue-equivalent phantoms (like CIRS) produce different HU from purely physical phantoms (like Catphan) at high densities, though both are valid. It underlines that whichever phantom is used for TPS calibration, the same should be used for consistency checks. If multiple phantoms are used interchangeably, one should be cautious when interpreting results; our solution could handle multiple reference baselines if needed.

One limitation of our study is that the CT scanner conditions remained stable (no tube

Article In Press

changes, etc.) during the evaluation period. Thus, we did not witness a scenario of a true calibration drift or failure to fully test the alert system. In the future, it would be valuable to intentionally adjust the scanner (e.g., apply a slight mis-calibration in the TPS) or use a different kV to see if the system correctly flags it. Another limitation is that we focused on a single energy (120kV), as that is predominantly used. For clinics that employ dual energies (for example, 140 kV for head protocols), the automation would need to be extended to handle separate calibrations for each energy, which is straightforward technically, but we did not report such data here. Additionally, our tolerance limits, while based on literature and reasonable, could be refined with more experience or stricter requirements if a clinic desires (for example, one could tighten water tolerance to ±20 HU if aiming for higher precision).

A practical consideration is that implementing this automation requires some programming expertise and QA system setup, which may be a barrier for some clinics. However, the code being open-source means users can adopt and modify it, and the integration with QATrack+ is well-documented by the QATrack+ community [23-25]. For centers without QATrack+, the script could simply output a report or log file, though the full benefits of trending and alerts might not be realized without a database.

We plan to expand the algorithm to incorporate additional phantoms as needed (for instance, the Gammex RMI phantom or others used in some clinics). Because our design is modular, recognizing a new phantom is a matter of adding its geometric template and known densities. We will also explore linking the QA results directly to the TPS calibration table. A possible future workflow is: if QATrack+ indicates a consistent bias (say all insert HU are ~+40 HU one month), the system could prompt the generation of a corrected HU-RED table and import it into the TPS after physicist review. This would create a closed feedback loop for maintaining calibration accuracy. While manual intervention is always needed for safety, automating the suggestion and calculation of a new table could save time and reduce errors. Another area for future research is applying the approach to CBCT systems. CBCT (cone-beam CT) in radiotherapy often has larger HU variability and is frequently not used for electron density directly, but with emerging techniques (HU-to-density for CBCT-based dose calculations [13]), it will become important. Our method, especially using the Catphan, could be extended to monthly CBCT HU checks and integrated into the same QA framework.

Finally, the success of automation here encourages similar efforts for other imaging QA tests. For example, automated analysis of low-contrast detectability, spatial resolution, etc., could be combined such that a single phantom scan yields a comprehensive QA report (some commercial solutions like RapidCHECK aim to do exactly this). In fact, RapidCHECK now offers an Advanced Electron Density Phantom with auto-analysis, which is conceptually similar to our approach. The difference is that our solution is fully open and customizable. RapidCHECK's introduction confirms the field's recognition of the need for automated CT calibration QA. Our work provides an independent verification that such automation is feasible and beneficial.

Our tolerances were partly based on ensuring any HU fluctuations remain dosimetrically insignificant. It is worth noting that modern TPS dose algorithms are robust to small density errors (through smoothing, etc.). Nakao et al. investigated tolerance levels of CT calibration and suggested that even a few percent error in electron density (which corresponds to tens of HU) is usually tolerable in photon dose calculations [27]. This underpins why ±50 HU in bone was considered acceptable in TG-66. Our results never approached those limits, suggesting that the scanner could even be stable enough to allow tighter tolerances. However, given that daily patient imaging has its own variability, there is a diminishing return to over-tightening QA thresholds. Our chosen values strike a balance and align with existing guidance.

One interesting observation is that the EasyCube phantom, being much smaller, did not significantly deviate in HU from the larger phantoms for soft tissues. Some might expect more beam hardening in a larger phantom to elevate soft-tissue HU slightly, but our data didn't show a strong effect in the 0-1.0 RED range. This suggests that the scanner's beam hardening correction is effective and perhaps calibrated assuming a body size. The bone differences are more due to composition than size. The advantage of the CIRS phantom is the number of tissue points (covering multiple bone densities, etc.), whereas Catphan's advantage is its multi-purpose nature. EasyCube, while limited in points, doubles as a dosimetry phantom. In practice, any can be used, but consistency in using the same type for QA is key. By integrating multiple phantoms into our program, we demonstrated that we could verify one against another (as an audit). This could be of use in external audits (a trial QA group could send a

Article in Press

phantom to verify a center's CT calibration; our software could analyze the scan and compare it to the center's baseline). In fact, Nakao et al. developed an audit phantom for CT calibration [27], indicating interest in inter-institution comparisons. Automation would facilitate such audits by ensuring each site analyzes the phantom in a standardized way.

The introduction of automated tools has the potential to transform QA practices; tedious tasks (and thus done rarely) can become routine. This fosters a more data-driven approach to QA. With abundant data points (monthly results), one can apply statistical process control, observe trends, and generally have greater insight into the machine's behavior. It moves QA from a reactive to a proactive stance. For example, one could calculate the mean and standard deviation of the monthly HU for each insert and set control limits at ±2σ; if the process starts to wander, action can be taken before an out-of-tolerance event occurs. In our case, the process variability was very low (σ on the order of a few HU), so our fixed.

CONCLUSION

The present study demonstrates that automated HU-RED calibration can provide stable and reproducible monthly QA results when implemented under standardized acquisition and analysis conditions. Although monthly consistency may also be influenced by scanner performance, phantom setup, and acquisition parameters, these factors were controlled throughout the study period. Under these conditions, the automated workflow improved reproducibility, reduced operator dependence, and supported efficient longitudinal monitoring of CT calibration stability.

ACKNOWLEDGMENT

The authors gratefully acknowledge the medical physics and radiotherapy staff of the International Hospital of Settat for their technical assistance with CT phantom acquisition and routine quality assurance measurements.

AUTHOR CONTRIBUTION

O. El Mouden conceived and designed the study, coordinated the project, acquired and analyzed the data, interpreted the findings, and drafted the manuscript. H. Sekkat contributed to methodological development, supervised data collection, validated the analyses, and critically revised the manuscript. A. Bannan participated in image acquisition, QA measurements, algorithm implementation, data verification, and contributed to the Methods section. R. El Baydaoui assisted with data acquisition, quality control procedures, and manuscript proofreading. M. Bougteb supported the experimental setup, detector calibration, and reviewed methodological components. M. Mkimel conducted the statistical analysis, prepared figures and tables, and contributed to the writing of the Results. A. Khallouqi provided technical support, equipment configuration, and participated in measurement sessions. M. Tahiri ensured data management and verification of clinical parameters and contributed to the final manuscript review. Y. Berrada offered scientific supervision, methodological guidance, and assisted in interpreting the clinical implications. F. Chaara provided expert clinical validation and critically reviewed the Discussion and Conclusion. O. El Rhazouani oversaw the project, ensured methodological accuracy, and granted final approval of the manuscript. All authors read and approved the final version of the work.

REFERENCES

1. Karl, J. G. Steel, and G. B. Warr, Phys. Eng. Sci. Med. **46** (2023) 1387.
2. E. K. Inness, V. Moutrie, and P. H. Charles, Australas. Phys. Eng. Sci. Med. **37** (2014) 385.
3. M. Nomura, S. Goto, M. Yoshioka *et al*., Phys. Eng. Sci. Med. **48** (2025) 1679.
4. A. T. Davis, S. Muscat, A. L. Palmer *et al*., BJR Open **1** (2019) 20190023.
5. E. Goerlitz, D. Albers, F. Cremers *et al*., *Influence of the CT Calibration Process on the Dose Computation in Radiotherapy Treatment Planning,* in: International Conference on Quality Assurance and New Techniques in Radiation Medicine, Book of Extended Synopses (2006) 226.
6. H. Sekkat, K. Elmansouri, A. Khallouqi *et al*., Radioprotection **59** (2024) 203.
7. H. Sekkat, K. Elmansouri, A. Khallouqi *et al*., Radiat. Eff. Defects Solids **179** (2024) 1295.
8. H. Sekkat, A. Khallouqi, O. El Rhazouani *et al*., Radiat. Eff. Defects Solids **180** (2025) 559.
9. A. M. Jaafar, H. Elsayed, M. M. Khalil *et al*., Prec. Radiat. Oncol. **6** (2022) 289.
10. H. Sekkat, A. Khallouqi, A. Bannan *et al*., Radiat. Phys. Chem. **237** (2025) 113068.

Article In Press

11. L. E. Lubis, I. Hariyati, D. Ryangga *et al.,* Atom Indones. **46** (2020) 69.
12. D. R. Putri, F. K. Hentihu, R. J. Stevenly *et al*., Atom Indones. **52** (2026) 105.
13. J. D. Saborido-Moral, M. Fernández-Patón, N. Tejedor-Aguilar *et al*., Phys. Med. **114** (2023) 103153.
14. Sun Nuclear, RapidCHECK™: Automated CT-to-density calibration & CT image quality analysis. https://www.sunnuclear.com/products/ rapidcheck-software. Retrieved in March (2025).
15. O. El Mouden, H. Sekkat, A. Bannan *et al*., Phys. Eng. Sci. Med. (2026).
16. O. El Mouden, H. Sekkat, A. Bannan *et al.*, Health Technol. **16** (2026) 743.
17. O. El Mouden, H. Sekkat, A. Bannan *et al*., Radiography **32** (2026) 103460.
18. S. Mutic, J. R. Palta, E. K. Butker *et al.,* Med. Phys. **30** (2003) 2762.
19. A. Saini, A. Singh, V. P. Pandey *et al*., ECS Trans. **107** (2022) 5729.
20. M. Nakao, S. Ozawa, H. Miura *et al*., Med. Phys. **47** (2020) 1509.
21. J. Annkah, I. Rosenberg, N. Hindocha *et al*., J. Med. Phys. **39** (2014) 133.
22. J. Kerns, *Pylinac: An Open-Source Quality Assurance Image Analysis Engine,* in: XXth International Conference on the use of Computers in Radiation therapy (ICCR) (2024).
23. C. Angers, R. Bottema, L. Buckley *et al.,* Health Phys. **117** (2019) 306.
24. L. Buckley, G. Heddon, I. Byrne *et al*., J. Med. Imaging Radiat. Sci. **51** (2020) 22.
25. I. Juma, L. Buckley, A. Thibeau *et al*., Med. Phys. **44** (2017) 4373.
26. S. Vatnitsky, Commissioning of Radiotherapy Treatment Planning Systems: Testing for Typical External Beam Treatment Techniques, International Atomic Energy Agency (IAEA), Vienna (2008).
27. M. Nakao, S. Ozawa, K. Yamada *et al*., J. Appl. Clin. Med. Phys. **19** (2018) 271.

Article In Press