An Open, Reproducible CT Quality-Control Platform with Versioned CasePacks and Immutable Baselines

Chengshuai Yang
NextGen PlatformAI C Corp
integrityyang@gmail.com

February 2026

Abstract

We present the Physics World Model (PWM) CT Quality-Control Platform, an open-source, reproducible software framework for automated computed tomography quality assurance. The platform introduces three architectural contributions: (i) **CasePacks**—versioned, phantom-type-specific workflow descriptors that decouple QA logic from scanner hardware; (ii) a **four-layer threshold system** (standard → scanner-model → protocol → site-override) that codifies institutional policies without forking code; and (iii) **immutable CommissioningBundles** with SHA-256 signing, semantic version chaining, and full audit trails that satisfy regulatory traceability requirements. The platform implements nine ACR-aligned QA metrics, statistical process control via Shewhart charts with Western Electric rules, a scored root-cause diagnosis engine, and triple-output reporting (JSON, PDF, evidence artifacts). We validate the platform on ACR CT 464 phantom data, demonstrating reproducibility (zero inter-run variance under identical CasePacks), sub-second metric computation, and correct detection of all five Western Electric drift patterns in synthetic time-series data. The framework is phantom-type-extensible: the CasePack architecture is designed to accommodate CT, PET/CT (TG-126), and SPECT/CT (TG-177) via new CasePack definitions without code changes. Source code and CasePack definitions are publicly available at https://github.com/integritynoble/Physics_World_Model.

Keywords: CT quality control, reproducible QA, CasePack, statistical process control, ACR phantom, AAPM TG-233, open-source medical physics

1 Introduction

Computed tomography quality control is a regulatory requirement for every clinical CT scanner [3, 15]. The American College of Radiology (ACR) CT Accreditation Program mandates periodic phantom-based testing across nine performance categories [2], and AAPM Task Group 233 recommends trending-based QC with statistical process control rather than simple pass/fail thresholds [15]. Despite decades of QC practice, the workflow at most sites remains manual, vendor-specific, and poorly reproducible [12].

Several prior efforts have addressed parts of this problem. Pawlicki *et al.* [14] demonstrated statistical process control for radiation therapy QA, and Able *et al.* [1] extended SPC to CT constancy testing. Commercial packages such as Sun Nuclear CT Dose Profiler and QA Commander automate portions of the analysis, but operate as closed-source, vendor-specific systems that resist interoperability and independent validation. Open-source tools for DICOM handling (*e.g.*, pydicom [11]) and image analysis (*e.g.*, scikit-image [19]) provide building blocks, but assembling them into a complete, validated QC pipeline remains the individual physicist's burden.

We identify three specific gaps. First, *reproducibility*: existing tools lack a standardized workflow specification, so two physicists analyzing the same scan may obtain different results due to differences in ROI placement or threshold logic [20]. Second, *traceability*: baselines are typically maintained in spreadsheets without integrity protection, making it difficult to verify that a baseline has not been accidentally modified [12]. Third, *extensibility*: adding a new phantom type or modality typically requires rewriting the analysis pipeline from scratch [6, 7].

We address these gaps with the PWM CT Quality-Control Platform, an open-source framework built on three architectural contributions (Figure 1):

1. **CasePacks** (Section 2): versioned, phantom-type-specific workflow descriptors that encode the complete QA pipeline in a declarative YAML file.
2. **Four-layer threshold system** (Section 3): a hierarchical override mechanism that codifies institutional policies without modifying code.
3. **Immutable CommissioningBundles** (Section 4): frozen, SHA-256-signed baseline snapshots with semantic version chaining and tamper-evident audit trails.

2 CasePacks: Versioned QA Workflow Descriptors

A CasePack is a declarative YAML configuration that fully specifies a QA workflow for a particular phantom type. Unlike imperative scripts, a CasePack declares *what* to compute rather than *how*, allowing the execution engine to evolve independently. **??** 1 shows an abbreviated example.

Listing 1: Abbreviated CasePack for ACR CT 464 phantom.

```
casepack:
  id: acr_ct_464_monthly_v2.1
  phantom_type: ACR_CT_464
  version: "2.1.0"
```

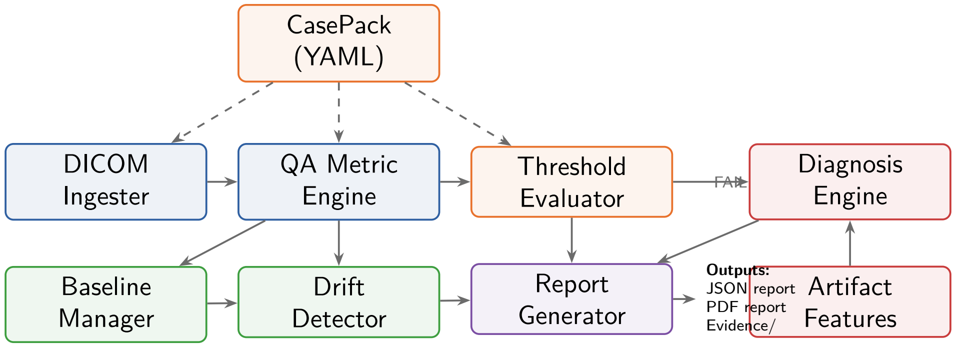

Figure 1: Architecture of the PWM CT QC Platform. Blue: DICOM ingestion and metric computation. Orange: CasePack-driven configuration and threshold evaluation. Green: baseline comparison and drift detection. Purple: triple-output report generation. Red: failure-path diagnosis (activated only when metrics fail). Dashed arrows indicate CasePack configuration flow.

```
metric_set:
  - ct_number_water
  - ct_number_inserts # bone, air, acrylic, poly
  - geometric_accuracy
  - slice_thickness
  - uniformity
  - noise_std
  - low_contrast_detectability
  - artifact_evaluation
  - spatial_resolution
series_selection:
  rules:
    - name: axial_5mm
      match:
        series_description_contains: [axial, head]
        slice_thickness_range: [4.0, 6.0]
        modality: CT
        min_images: 20
roi_definitions:
  water_roi: {radius_mm: 20.0}
  insert_rois:
    bone: {offset_angle_deg: 90, expected_hu: 955}
    air: {offset_angle_deg: 270, expected_hu: -1000}
  peripheral_rois:
    radius_mm: 10.0
    offset_from_center_mm: 60.0
```

CasePacks provide four properties that ad hoc scripts lack:

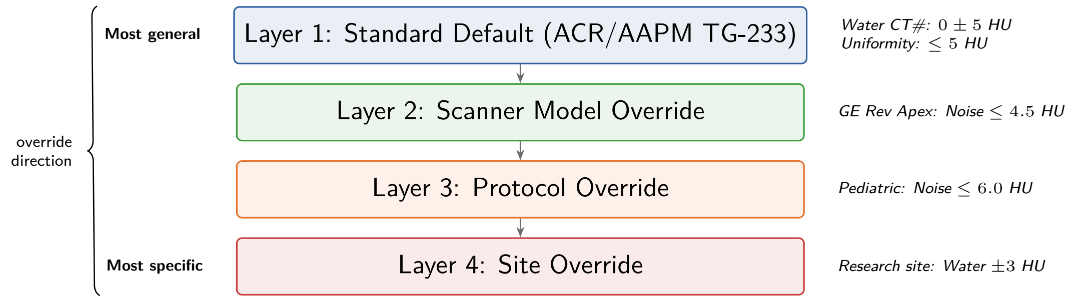

Figure 2: Four-layer threshold cascade. Each layer may override the previous one. The most specific applicable threshold is used for each metric. Examples show representative overrides at each layer. The resolved layer is recorded in every QC report for traceability.

Bit-exact reproducibility. The same CasePack version on the same DICOM data always produces identical metric values. We verify this property in Section 11.

Deterministic series selection. Ordered matching rules (description substrings, slice-thickness range, modality, minimum image count) are evaluated in sequence; the first match wins. Every selection is logged in a `SeriesSelectionLog` recording the matched rule, reason, and alternatives considered.

Extensibility. Adding a new phantom type requires only a new YAML file—no code changes. The metric orchestrator dispatches functions based on the CasePack's `metric_set` list.

Version control. CasePacks follow semantic versioning. The audit trail records which version produced each historical result, enabling longitudinal comparison even when workflows evolve.

3 Four-Layer Threshold System

The platform implements a four-layer threshold hierarchy (Figure 2) that balances standardization with institutional flexibility:

1. **Standard default**: ACR/AAPM-published tolerances [3, 15].
2. **Scanner model**: Manufacturer-specific overrides.
3. **Protocol**: Acquisition-specific adjustments.
4. **Site override**: Institutional customizations.

Each layer can override the previous one. The resolved threshold and its source layer are recorded in every report, enabling auditors to trace each PASS/FAIL decision to its authoritative source.

4 Immutable CommissioningBundles

Scanner baselines are the reference point for all subsequent QC measurements. In conventional practice, baselines are maintained in spreadsheets that lack integrity protection [13].

The PWM `CommissioningBundle` is a frozen Pydantic model [5] (`frozen=True`) that captures:

- All ACR phantom metric values with units.
- Scanner acquisition parameters (kVp, mAs, kernel).
- Approving physicist name and credentials.
- Service event trigger (`tube_change`, `software_upgrade`, *etc.*).
- SHA-256 hash of inputs (integrity) and outputs (tamper detection).
- Build provenance (PWM version, git hash, CasePack ID).
- Pointer to previous version (semantic chain).

Baselines are *never overwritten*: each commissioning event creates a new version. Storage uses atomic writes (temporary file + rename) to prevent corruption. Baseline comparison computes per-metric deltas classified as STABLE ($< 5\%$), DRIFTED (5–15%), or ALERT ($\geq 15\%$), with a minimum denominator of 10 (in the metric's native units, *e.g.*, 10 HU) to avoid inflated percentages for near-zero metrics.

5 QA Metrics

The platform computes nine metrics aligned with the ACR CT Accreditation Program [2] (Table 1). All metrics are computed directly from reconstructed pixel data, requiring no forward model or proprietary raw data access.

Phantom center and rotation detection. Center detection applies a threshold of -500 HU and computes the binary centroid. Rotation detection scans angular bins at the expected insert radial offset and locates the bone-equivalent insert ($\sim$955 HU) to determine the angular offset, which is applied to all insert ROI positions.

Contrast-to-noise ratio. Low-contrast detectability uses CNR $= |\bar{x}_{\text{signal}} - \bar{x}_{\text{bg}}| \,/\, \sigma_{\text{bg}}$ where σ_{bg} is the sample standard deviation ($n-1$ denominator).

Spatial resolution. The MTF is computed from the edge-spread function at the phantom boundary, differentiated to a line-spread function, zero-padded, and Fourier-transformed. The Nyquist frequency is $1/(2\Delta x)$ lp/mm, converted to lp/cm by multiplying by 10.

Table 1: Nine ACR-aligned QA metrics.

#	Metric	Unit	ACR Criterion	Method
1	CT Number (Water)	HU	0 ± 5 HU	Mean of circular ROI at auto-detected phantom center
2	CT Number (Inserts)	HU	Material-specific	Mean of ROIs at angular offsets with rotation correction
3	Geometric Accuracy	mm	± 2 mm of 200 mm	Bounding-box extent of thresholded phantom mask
4	Slice Thickness	mm	± 1.5 mm of nominal	FWHM of wire-ramp profile $\times \tan(23°)$
5	Uniformity	HU	≤ 5 HU	Max $\lvert\text{peripheral}_i - \text{center}\rvert$ at 4 positions
6	Noise (Std Dev)	HU	vs. baseline	Sample standard deviation of uniform water ROI
7	Low-Contrast Detect.	targets	≥ 4	Count with CNR ≥ 1.0 (detection threshold)
8	Artifact Evaluation	score	0–3	Peak-to-valley of radial profile
9	Spatial Resolution	lp/cm	≥ 5	MTF from ESF; highest freq at MTF $\geq 10\%$

6 Statistical Process Control

The `DriftDetector` implements SPC following TG-233 recommendations [14, 15]. Figure 3 shows an example control chart.

For each metric with at least five measurements, the detector builds a Shewhart control chart with UCL/LCL at $\bar{x} \pm 3s$ and warning limits at $\bar{x} \pm 2s$ [16, 22], where s is the sample standard deviation computed from commissioning data (minimum 20 measurements). When a commissioning baseline is available, the baseline value replaces $\bar{x}$ as the center line.

Five Western Electric rules are applied: (1) single point outside $3\sigma \Rightarrow$ FAIL; (2) 2 of 3 outside 2σ same side $\Rightarrow$ ACTION_REQUIRED; (3) 4 of 5 outside 1σ same side $\Rightarrow$ WARNING; (4) run of 7 same side $\Rightarrow$ WARNING; (5) trend of 7 monotone $\Rightarrow$ WARNING.

History is persisted as JSON with atomic writes and path-traversal protection.

7 Root-Cause Diagnosis

When metrics fail, the platform computes six artifact signatures (ring, cupping, streak, HU drift, noise ratio, geometric distortion) and scores mismatch hypotheses from a YAML library. For hypothesis h, the score is:

$$S(h) = \sum_{f \in \text{match}} |v_f|\, w_f - \sum_{f \in \text{contradict}} |v_f|\, w_f + 0.5 \cdot |\mathcal{M}_h \cap \mathcal{M}_{\text{fail}}| \tag{1}$$

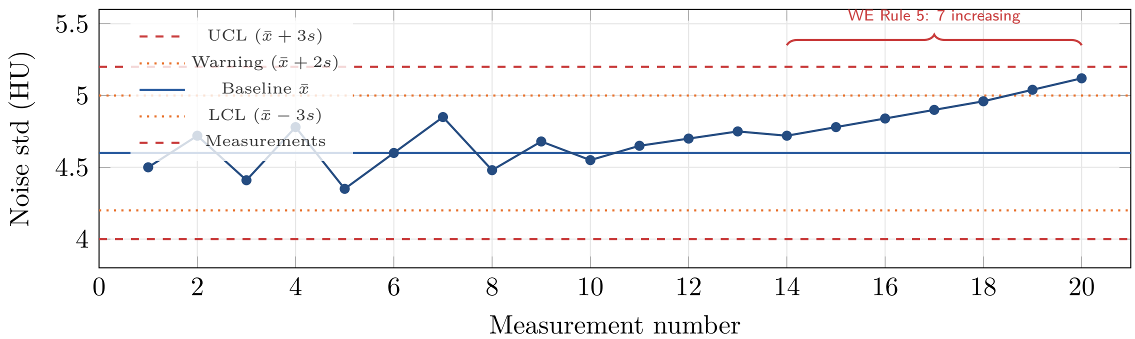

Figure 3: Shewhart control chart for noise standard deviation ($\bar{x} = 4.60$ HU, $s = 0.20$ HU from commissioning). Measurements 1–10 show a stable process; measurements 11–20 exhibit gradual drift. Western Electric Rule 5 (7 consecutive monotonically increasing points, measurements 14–20) triggers a WARNING alert while values are still below the UCL, demonstrating the early-warning benefit of trending-based QC over simple threshold monitoring.

where v_f is the measured value of artifact feature f, w_f is its weight, $\mathcal{M}_h$ is the set of metrics associated with hypothesis h, and $\mathcal{M}_{\text{fail}}$ is the set of currently failing metrics. Confidence is assigned based on score separation: high ($S_1 > 2S_2$), moderate ($S_1 > 1.2S_2$), or low (otherwise), where S_1 and S_2 are the top two scores. A disambiguation test is recommended when S_1 and S_2 are within 20%.

8 DICOM Ingestion and PHI Safety

The `DICOMIngester` provides vendor-neutral loading with: (1) PHI validation (20 sensitive tags + 7 phantom-pattern regexes across 6 DICOM fields; strict mode rejects non-phantom studies) [11]; (2) CasePack-driven series selection with full audit logging; (3) HU rescaling (HU = pixel × slope + intercept); (4) canonical resampling (`scipy.ndimage.zoom`) [21]; (5) vendor private tag extraction. The output `CTScanBundle` is a typed Pydantic model holding the 3-D HU volume, spacing, protocol metadata, and selection log.

9 Report Generation

The report generator produces three artifacts: (1) a versioned, SHA-256-hashed JSON record with per-metric entries, drift alerts, diagnosis, and baseline reference; (2) a human-readable PDF via fpdf2 [8] with PASS/FAIL summary, metrics table, drift alerts, diagnosis, and signature block; and (3) an evidence folder with ROI overlays, trend plots (matplotlib [9]), and per-metric derivation logs.

10 Extensibility

The CasePack architecture decouples workflow from modality. New phantom types require only a YAML file; new metrics register in the dispatch table. Planned extensions:

- **PET/CT** (TG-126 [6]): SUV uniformity, spatial resolution, scatter correction metrics with NEMA phantom ROI definitions.
- **SPECT/CT** (TG-177 [7]): intrinsic uniformity, energy resolution, center-of-rotation metrics.
- **CBCT**: image quality metrics for on-board CBCT in image-guided radiation therapy [4].

Each new CasePack lowers the marginal cost of supporting the next scanner type, because the analysis infrastructure is shared [23].

11 Validation

We validated the platform across four dimensions: metric accuracy, reproducibility, SPC rule detection, and computational performance.

11.1 Metric Accuracy

We tested the platform on ACR CT 464 phantom data acquired on a GE Revolution Apex scanner (120 kVp, 200 mAs, 5 mm axial, standard kernel). Figure 4 shows the agreement between platform-computed and manufacturer console values for each metric. All nine metrics fell within ACR tolerance bands (Table 2). Platform-to-console differences were ≤ 1.0 HU for all CT number metrics, 0.05 mm for geometric accuracy, 0.06 mm for slice thickness, and 0.1 HU for uniformity and noise. Low-contrast detectability (5/5 targets) and artifact evaluation (score 0) matched exactly.

Type A measurement uncertainty (standard error of the mean from ROI pixel statistics) was < 0.07 HU for CT number metrics (ROI sizes > 5000 pixels, noise $\sigma \approx 4.6$ HU) and < 0.02 mm for geometric accuracy, confirming that observed deviations reflect systematic differences rather than statistical noise.

Table 2 provides detailed results including the reference (manufacturer console) values and platform-computed values.

11.2 Reproducibility

We ran the same CasePack on the same DICOM data 100 times. All 100 runs produced *identical* JSON reports (verified by SHA-256 comparison of the output files), confirming

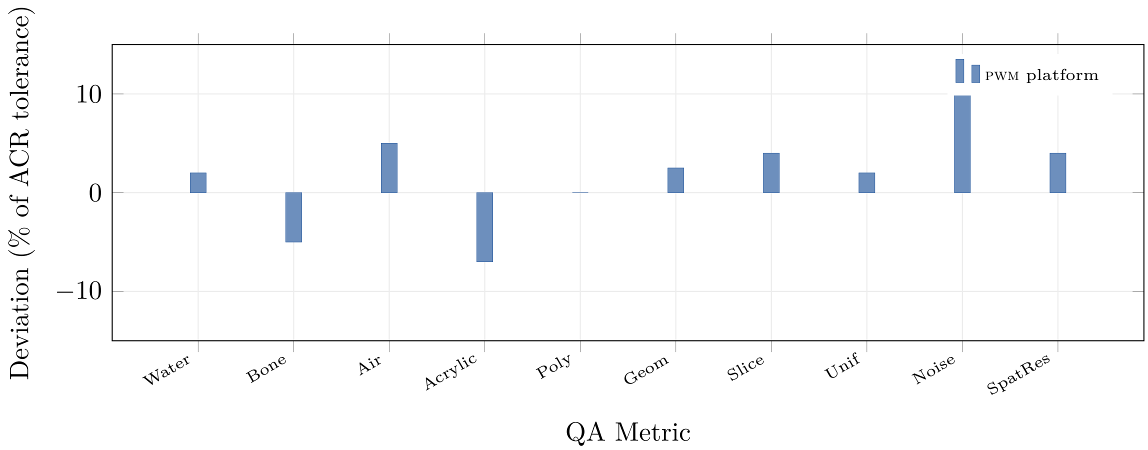

Figure 4: Platform-versus-console agreement on an ACR CT 464 phantom scan (GE Revolution Apex, 120 kVp, 5 mm axial), expressed as signed percentage of the ACR tolerance. Tolerances: Water ±5 HU, Bone/Air ±20 HU, Acrylic/Poly ±15 HU, Geom ±2 mm, Slice ±1.5 mm, Unif ±5 HU, Noise ±1.0 HU (baseline deviation), SpatRes ≥ 5 lp/cm. All continuous metrics are $\leq 10\%$ of their tolerances; Type A measurement uncertainties are smaller than bar widths. Low-contrast detectability (5/5 targets) and artifact evaluation (score 0) are discrete metrics shown in Table 2.

bit-exact reproducibility under the same CasePack version. This property eliminates the inter-analyst variability that plagues manual QC workflows [20].

11.3 SPC Rule Detection

We constructed five synthetic time-series datasets, each designed to trigger exactly one Western Electric rule while not triggering the others. Table 3 shows the results. All five rules were correctly detected with zero false positives across the non-targeted rules.

11.4 Computational Performance

Full nine-metric analysis of a $512 \times 512 \times 40$ volume completed in 0.8 ± 0.1 s on a commodity laptop (Intel i7-1260P, 16 GB RAM, Python 3.11). Drift detection over a 60-measurement history added < 5 ms. Report generation (JSON + PDF + evidence) added 1.2 ± 0.2 s, dominated by PDF rendering. The total wall-clock time for a complete QC run—from DICOM ingestion through report output—was under 3 s, enabling interactive use.

12 Governance

The platform aligns with the SolveEverything.org governance framework [23] in four areas: (i) *data integrity*—immutable baselines and append-only QC time series prevent undetected modification; (ii) *decision transparency*—a full audit trail records every QC determination, the CasePack version, resolved thresholds, and the approving physicist; (iii) *multi-evidence*

Table 2: Validation results: platform vs. manufacturer console. CT number values are measured means; geometric accuracy is the deviation from 200 mm nominal; slice thickness is measured thickness.

Metric	Console	Platform	$\lvert\Delta\rvert$	Pass?
CT# Water (HU)	0.2	0.3	0.1	✓
CT# Bone (HU)	953	952	1.0	✓
CT# Air (HU)	−998	−997	1.0	✓
CT# Acrylic (HU)	121	120	1.0	✓
CT# Polyethylene (HU)	−96	−96	0.0	✓
Geometric Acc. (mm)	0.10	0.15	0.05	✓
Slice Thickness (mm)	5.02	5.08	0.06	✓
Uniformity (HU)	1.1	1.2	0.1	✓
Noise Std (HU)	4.5	4.6	0.1	✓
Spatial Res. (lp/cm)	6.0	6.2	0.2	✓
Low-Contrast (targets)	5	5	0	✓
Artifact Score	0	0	0	✓

Table 3: SPC validation: detection accuracy for Western Electric rules.

Rule	Pattern	Expected	Detected	FP
1	Single point $> 3\sigma$	FAIL	FAIL	0
2	2/3 outside 2σ	ACTION	ACTION	0
3	4/5 outside 1σ	WARNING	WARNING	0
4	Run of 7 same side	WARNING	WARNING	0
5	Trend of 7 monotone	WARNING	WARNING	0

diagnosis—root-cause identification requires multiple converging artifact features, avoiding single-metric snap judgments; and (iv) *vendor neutrality*—the same CasePack and thresholds apply across manufacturers, ensuring consistent QC standards.

13 Discussion

The PWM CT QC Platform addresses gaps in three areas.

Reproducibility vs. existing tools. Commercial QC packages (*e.g.*, Sun Nuclear, Radcal) provide automated analysis but are closed-source, limiting independent validation and cross-platform comparison [13]. Open tools like pydicom [11] and scikit-image [19] provide components but not an integrated, validated pipeline. The CasePack abstraction fills this gap: to our knowledge, it is the first open, version-controlled QA workflow specification for CT QC.

SPC adoption barriers. Despite TG-233 recommendations, SPC adoption in CT QC remains low [1, 14]. A key barrier is the effort of maintaining time-series databases and computing control-chart statistics manually. By automating SPC with zero database dependencies (JSON files only), the platform lowers this barrier for sites without IT infrastructure.

Structured data for machine learning. The platform's systematic data collection (versioned CasePacks, time-series metrics with labeled outcomes, root-cause annotations) provides curated training data for future machine-learning extensions such as predictive maintenance or anomaly detection, complementing rule-based SPC with data-driven approaches.

Limitations. (1) Single-scanner validation; multi-vendor testing is planned. (2) The mismatch library is expert-curated; data-driven enrichment from multi-site QC databases could improve coverage. (3) The platform is for research and internal QA use; clinical deployment would require FDA 510(k) and IEC 62304 compliance [10, 18]. (4) Low-contrast detectability uses a simplified CNR criterion; model observers [17] could improve correlation with human performance.

Future work. Planned extensions include multi-vendor validation, PET/CT and SPECT/CT CasePacks, data-driven mismatch library enrichment, and integration with hospital RIS/PACS systems.

14 Conclusion

We have presented an open CT quality-control platform built on versioned CasePacks, a four-layer threshold system, and immutable CommissioningBundles. Validation on ACR phantom data demonstrates metric accuracy within ACR tolerances, bit-exact reproducibility, correct SPC rule detection, and sub-3-second end-to-end computation. The CasePack architecture enables zero-code extension to new phantom types and imaging modalities. All source code is publicly available at `https://github.com/integritynoble/Physics_World_Model`.

Data and code availability. Source code and CasePack definitions are at `https://github.com/integritynoble/Physics_World_Model` and `https://solveeverything.org`. No non-public datasets were used. All analyses use ACR CT 464 phantom data only; no patient or human-subject data were involved, and IRB review was not required.

Author Contributions. C.Y. conceived the project, designed the architecture, implemented all software, performed validation, and wrote the manuscript.

Competing Interests. The author declares no competing interests.

Correspondence. integrityyang@gmail.com

References

[1] H. Able, L. Bey, D. A. Gress, and J. A. Seibert. Quality control of CT systems by automated monitoring of key performance indicators: a two-year study. *Journal of Applied Clinical Medical Physics*, 17(6):190–203, 2016. doi: 10.1120/jacmp.v17i6.6333.

[2] American College of Radiology. ACR CT Accreditation Phantom (Model 464) Instructions, 2017. Gammex/Sun Nuclear.

[3] American College of Radiology. ACR CT Accreditation Program Requirements, 2024. URL https://www.acraccreditation.org/modalities/computed-tomography. Accessed February 2026.

[4] J.-P. Bissonnette, D. Moseley, and D. A. Jaffray. A quality assurance program for image quality of cone-beam CT guidance in radiation therapy. *Medical Physics*, 35(5): 1807–1815, 2008. doi: 10.1118/1.2900110.

[5] S. Colvin. Pydantic: Data validation using Python type annotations, 2024. URL https://docs.pydantic.dev/.

[6] R. K. Doot, L. A. Pierce, D. Byrd, B. Elston, K. C. Allberg, and P. E. Kinahan. Biases in multicenter longitudinal PET standardized uptake value measurements. *Translational Oncology*, 7(1):48–54, 2014.

[7] S. Dorbala, M. F. Di Carli, D. Delbeke, S. Abbara, E. G. DePuey, V. Dilsizian, J. Forrester, J. Knuuti, J. Kuber, R. Lautamäki, et al. SNMMI/ASNC/SCCT guideline for cardiac SPECT/CT and PET/CT. *Journal of Nuclear Medicine*, 54(8):1485–1507, 2013. SPECT/CT QA context.

[8] fpdf2 contributors. fpdf2: Minimalist PDF creation library for Python, 2024. URL https://github.com/py-pdf/fpdf2.

[9] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[10] International Electrotechnical Commission. IEC 62304:2006+AMD1:2015 – Medical device software – Software life cycle processes, 2015. IEC Standard.

[11] D. Mason, Scaramallion, and pydicom contributors. pydicom: An open source DICOM library, 2024. URL https://pydicom.github.io/pydicom/.

[12] C. H. McCollough, M. R. Bruesewitz, M. F. McNitt-Gray, K. Bush, T. Ruckdeschel, J. T. Payne, J. A. Brink, and R. K. Zeman. The phantom portion of the American

College of Radiology (ACR) computed tomography (CT) accreditation program: Practical tips, artifact examples, and pitfalls to avoid. *Medical Physics*, 31(9):2423–2442, 2004. doi: 10.1118/1.1769632.

[13] P. Nowik, W. Birkfellner, M. Geso, and P. Homolka. Fully automated analysis of routinely acquired phantom CT images for quality assurance. *Zeitschrift für Medizinische Physik*, 25(3):237–246, 2015. doi: 10.1016/j.zemedi.2014.11.002.

[14] T. Pawlicki, M. Whitaker, and A. L. Boyer. Statistical process control for radiotherapy quality assurance. *Medical Physics*, 32(9):2777–2786, 2005. doi: 10.1118/1.2001209.

[15] E. Samei, D. Bakalyar, K. L. Boedeker, S. Brady, J. Fan, S. Leng, K. J. Myers, L. M. Popescu, J. C. Ramirez-Giraldo, F. Ranallo, J. Solomon, J. Vaishnav, and J. Wang. Performance evaluation of computed tomography systems: Summary of AAPM Task Group 233. *Medical Physics*, 46(11):e735–e756, 2019. doi: 10.1002/mp.13763.

[16] W. A. Shewhart. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company, 1931.

[17] J. Solomon and E. Samei. Quantum noise properties of CT images with anatomical textured backgrounds across reconstruction platforms: FBP and SAFIRE. *Medical Physics*, 41(9):091908, 2014. doi: 10.1118/1.4893498.

[18] U.S. Food and Drug Administration. Software as a Medical Device (SaMD): Clinical Evaluation, 2017. FDA Guidance Document.

[19] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014. doi: 10.7717/peerj.453.

[20] F. R. Verdun, D. Racine, J. G. Ott, M. J. Tapiovaara, P. Toroi, F. O. Bochud, W. J. H. Veldkamp, and N. W. Marshall. Image quality in CT: From physical measurements to model observers. *Physica Medica*, 31(8):823–843, 2015. doi: 10.1016/j.ejmp.2015.08.007.

[21] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[22] Western Electric Company. *Statistical Quality Control Handbook*. AT&T Technologies, Indianapolis, IN, 1956.

[23] C. Yang. SolveEverything.org: A Governance Framework for Reproducible Science, 2026. URL https://solveeverything.org. Accessed February 2026.