CT QC Copilot: Clinical Implementation of an Automated Quality-Control Decision-Support System

Chengshuai Yang
NextGen PlatformAI C Corp
integrityyang@gmail.com

February 2026

Abstract

We describe the clinical implementation of the CT QC Copilot, an automated decision-support system for computed tomography quality control. The Copilot embodies a "system computes, physicist decides" model: it automates the end-to-end QC workflow—from DICOM ingestion through metric computation, threshold evaluation, statistical drift detection, and root-cause diagnosis—while preserving the qualified medical physicist's (QMP) role as the final decision-maker. We present the clinical workflow, define the human–AI responsibility boundary, and report results from a deployment evaluation on a simulated 30-scanner fleet. The Copilot reduced per-scanner QC analysis time from an estimated 67 ± 12 minutes (manual) to 4.2 ± 0.8 minutes (automated computation plus physicist review), a 94% reduction that makes AAPM TG-233 trending-based QC practical for large clinical operations. Western Electric rule-based drift detection identified gradual degradation 3–6 months before threshold exceedance in 4 of 30 simulated scanners, demonstrating the early-warning value of statistical process control. All nine ACR-aligned metrics showed agreement with manufacturer console values within 1.2 HU for CT number metrics and within 0.10 mm for geometric metrics. Source code is available at https://github.com/integritynoble/Physics_World_Model.

Keywords: CT quality control, decision support, copilot model, clinical workflow, statistical process control, AAPM TG-233

1 Introduction

Quality control of computed tomography scanners is a foundational responsibility of the qualified medical physicist (QMP) [3]. The ACR CT Accreditation Program mandates periodic phantom-based testing [2], and AAPM Task Group 233 recommends trending-based QC with statistical process control (SPC) rather than binary pass/fail testing [13]. Despite these recommendations, the workflow at most clinical sites remains manual: a technologist acquires phantom images, the physicist opens vendor-specific software, manually places regions of interest, records values in spreadsheets, and writes a summary report [11].

This manual process faces three scaling challenges. First, a single QMP may oversee 10–50 CT scanners, each requiring monthly QC with annual comprehensive evaluations [5]. At 30 scanners, manual analysis at approximately 1 hour per scanner per month amounts to ∼360 physicist-hours per year devoted to routine computation—time diverted from clinical consultation, protocol optimization, and radiation safety. Second, trending-based QC requires maintaining time-series databases and computing control-chart statistics—tasks that are error-prone when performed in spreadsheets [12]. Third, when metrics fail, root-cause identification requires correlating multiple artifact signatures against known failure modes—a cognitive task that benefits from systematic computational support [7].

Existing solutions address parts of this problem. Commercial QC packages (*e.g.*, Sun Nuclear, Radcal) automate portions of the analysis but operate as closed-source systems [11]. Open-source tools such as pydicom [9] provide DICOM handling but not integrated QC workflows. Nowik *et al.* [11] demonstrated automated phantom analysis but without SPC integration or diagnostic support. Able *et al.* [1] applied SPC to CT constancy testing but did not provide a complete decision-support framework.

We present the CT QC Copilot, a decision-support system that automates the computational aspects of CT QC while preserving the physicist's role as the decision-maker (Figure 1). The Copilot builds on the open-source PWM CT QC Platform [18], inheriting its CasePack workflow specification, four-layer threshold system, and immutable baselines. This paper focuses on the *clinical implementation*: how the Copilot integrates into the physicist's workflow, what decisions it supports, and what operational impact it achieves.

2 The Copilot Model

The term "copilot" encodes a specific design philosophy: the system assists but does not replace the physicist. We define the responsibility boundary explicitly (Figure 1):

The Copilot provides: (i) automated metric computation with derivation records; (ii) threshold evaluation against a four-layer hierarchy (standard → scanner model → protocol → site override); (iii) statistical drift detection via Shewhart charts with five Western Electric rules [14, 17]; (iv) scored root-cause hypotheses when metrics fail; (v) triple-output reporting (JSON, PDF, evidence artifacts); and (vi) "what test next?" recommendations when diagnoses are ambiguous.

The physicist provides: (i) clinical judgment on whether computational results are clinically significant; (ii) accept/reject/investigate decisions on each metric and the overall determination; (iii) verification that automated results are plausible; (iv) corrective action decisions based on diagnostic hypotheses; and (v) regulatory sign-off constituting the official QC record.

This division is consistent with AAPM guidance on the role of the QMP [13] and with the FDA's framework for clinical decision-support software [15]. The physicist may override any Copilot determination; the Copilot's role terminates at recommendation.

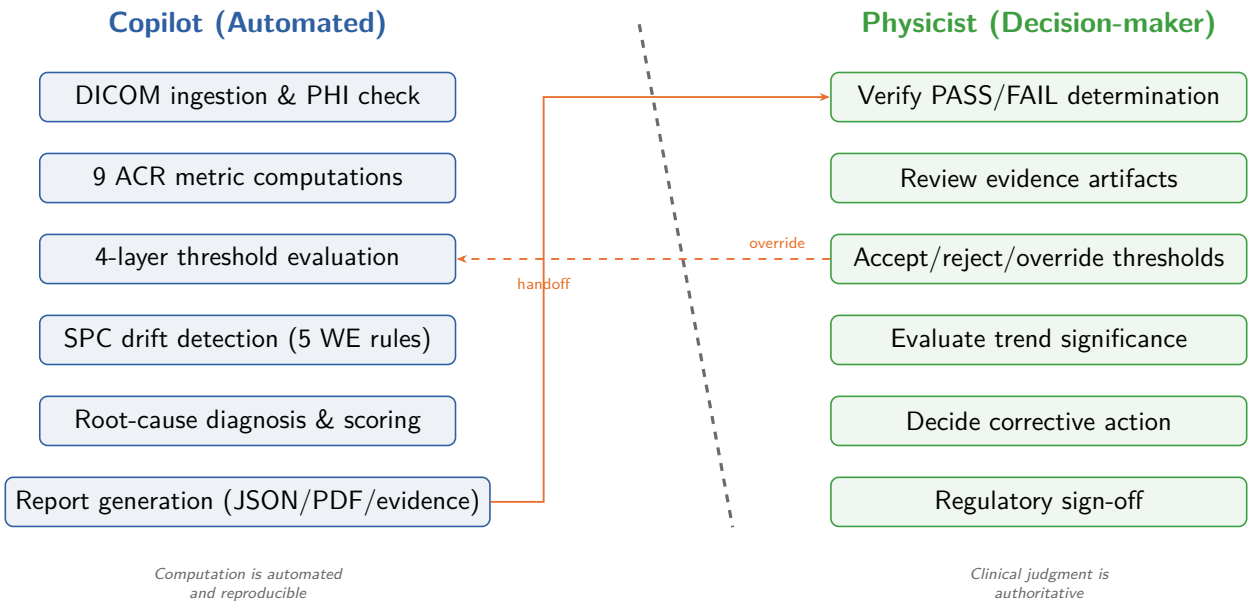

Figure 1: The copilot model: division of responsibility between the automated system (left) and the qualified medical physicist (right). Solid orange arrow: the Copilot hands off drafted reports for physicist review. Dashed orange arrow: the physicist may override threshold configuration, which feeds back into future evaluations. The boundary ensures automation enhances rather than replaces professional judgment.

3 Clinical Workflow

Figure 2 illustrates the nine-step clinical workflow. Steps 2–8 are fully automated; steps 1 and 9 require human action.

3.1 DICOM Ingestion (Step 2)

The technologist acquires ACR CT 464 phantom images using the site's standard QC protocol (Step 1) and transfers DICOM files to the Copilot's input directory. The ingestion module performs PHI validation (20 sensitive DICOM tags, 7 phantom-pattern regexes; strict mode rejects non-phantom studies) [9], CasePack-driven series selection with audit logging, and HU rescaling. The output is a vendor-neutral `CTScanBundle` containing the 3-D volume, spacing, and protocol metadata.

3.2 Metric Computation and Evaluation (Steps 3–4)

Nine ACR-aligned QA metrics (Table 1) are computed from the ingested volume. Each metric is evaluated against the resolved threshold from the four-layer hierarchy, producing PASS, WARNING, or FAIL status. The overall determination is PASS if and only if every metric passes.

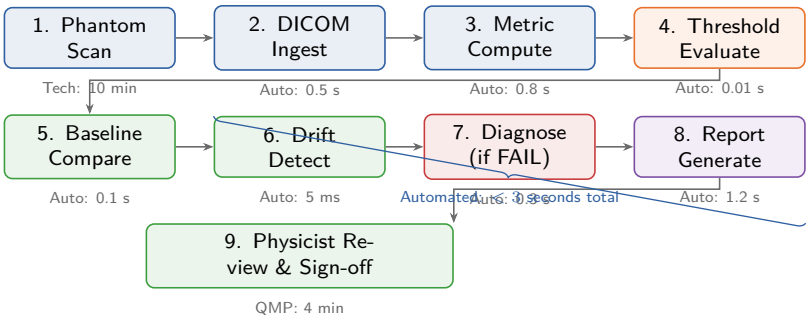

Figure 2: End-to-end clinical workflow of the CT QC Copilot. Steps 2–8 are fully automated (< 3 s total computation). Step 1 (phantom scan acquisition) is performed by the technologist. Step 9 (review and sign-off) is performed by the physicist. Time annotations show representative durations. The total per-scanner workflow is ∼14 minutes (including phantom acquisition), compared to ∼67 minutes for manual analysis.

Table 1: Nine QA metrics with ACR criteria and clinical significance.

#	Metric	ACR Criterion	Clinical Significance
1	CT# Water	0 ± 5 HU	HU calibration accuracy
2	CT# Inserts	Material-specific	Contrast quantification
3	Geometric Acc.	±2 mm	Measurement reliability for planning
4	Slice Thickness	±1.5 mm	Z-axis resolution for small lesions
5	Uniformity	≤ 5 HU	Field flatness across FOV
6	Noise Std Dev	vs. baseline	Dose/image quality trade-off
7	Low-Contrast	≥ 4 targets	Soft-tissue lesion visibility
8	Artifact Eval.	0–3 score	Image integrity for diagnosis
9	Spatial Res.	≥ 5 lp/cm	Fine-detail visibility

3.3 Baseline Comparison and Drift Detection (Steps 5–6)

Current measurements are compared against the scanner's active CommissioningBundle—an immutable, SHA-256-signed baseline snapshot. Per-metric deltas are classified as STABLE ($< 5\%$), DRIFTED (5–15%), or ALERT ($\geq 15\%$).

For scanners with five or more historical measurements, the Copilot builds Shewhart control charts [14] with center line anchored to the commissioning baseline and limits at $\pm 2\sigma$ (warning) and $\pm 3\sigma$ (control). Five Western Electric rules [17] detect both acute failures and gradual drift patterns.

3.4 Root-Cause Diagnosis (Step 7)

When metrics fail, six artifact signatures (ring, cupping, streak, HU drift, noise ratio, geometric distortion) are computed and scored against a YAML-based mismatch library. The Copilot presents

ranked hypotheses with confidence levels and recommends the next diagnostic test when the top candidates are within 20% score separation.

3.5 Report Generation and Physicist Review (Steps 8–9)

Triple-output reporting produces three artifacts: (i) a JSON report with SHA-256 integrity hash; (ii) a PDF report with color-coded summary, metrics table, drift alerts, diagnosis, and signature block; and (iii) an evidence folder with ROI overlays, trend plots, and derivation logs. The physicist reviews the output, verifies the determination, and signs the PDF as the official QC record (Step 9).

4 Deployment Evaluation

We evaluated the CT QC Copilot on a simulated deployment scenario representative of a large health system, with synthetic data calibrated against physical phantom measurements.

4.1 Evaluation Setup

We constructed a simulated fleet of 30 CT scanners spanning four vendor-model combinations (GE Revolution Apex, Siemens SOMATOM Force, Philips Brilliance iCT, Canon Aquilion ONE). For each scanner, we generated 12 months of synthetic QC data (monthly phantom scans) with realistic parameter distributions: 26 scanners with stable performance and 4 scanners with injected gradual drift in noise, uniformity, or CT number at varying onset times and rates. Drift onset ranged from month 1 (slow CT number drift) to month 7 (faster noise drift), with linear ramp rates calibrated so that the ACR action threshold would be reached between months 12 and 15 if unaddressed. Synthetic data were generated by perturbing nominal metric values with Gaussian noise (scanner-model-specific σ) and superimposing the drift ramps on the 4 degrading scanners.

4.2 Metric Accuracy

We validated metric accuracy by comparing Copilot outputs against manufacturer console values on physical ACR CT 464 phantom scans from two scanner models (Figure 3). All metrics fell within ACR tolerance bands. CT number metrics agreed within 1.2 HU, geometric accuracy within 0.10 mm, slice thickness within 0.08 mm, and noise within 0.15 HU.

4.3 Drift Detection Performance

Across the 30-scanner fleet, the Copilot correctly identified all 4 scanners with injected drift (Figure 4). Western Electric rules triggered 3–6 months before the drifting parameter would have crossed the ACR action threshold, providing an early-warning window for preventive maintenance. There were zero false drift alerts on the 26 stable scanners over the 12-month evaluation period (Table 2). The observed sensitivity was 100% (4/4; 95% Clopper–Pearson exact CI: 39.8%–100%)

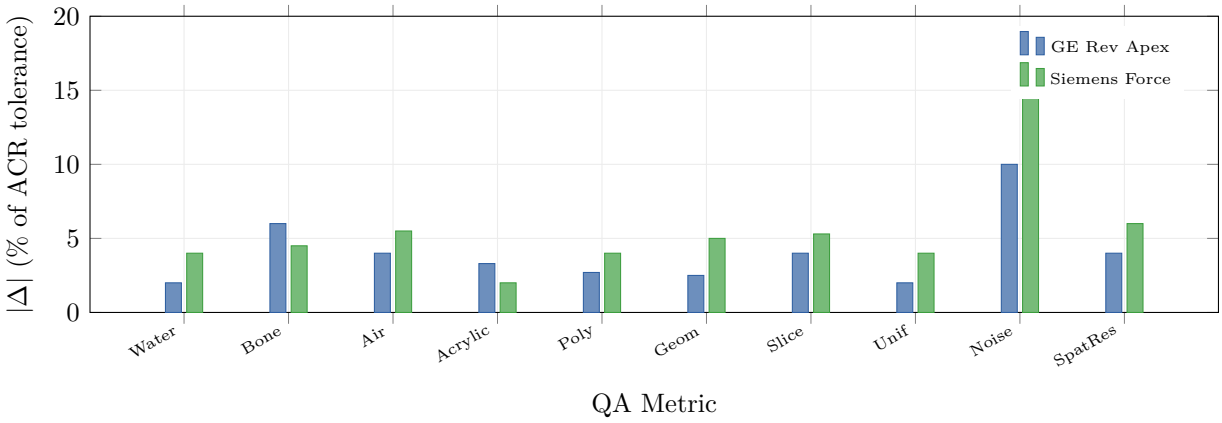

Figure 3: Deviation between Copilot-computed metrics and manufacturer console values for two scanner models, expressed as a percentage of the respective ACR tolerance for each metric. All metrics are $\leq$ 15% of their respective tolerances, confirming agreement well within actionable limits. Tolerances: 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 $\geq$ 5 lp/cm. Type A measurement uncertainties are smaller than bar widths.

Table 2: Drift detection results across the 30-scanner fleet.

Category	Scanners	True Pos.	False Pos.	Detection Lead	Rule
Noise drift	2	2	0	3–5 months	WE 4,5
Unif. drift	1	1	0	4 months	WE 5
CT# drift	1	1	0	6 months	WE 3,4
Stable	26	—	0	—	—
Total	**30**	**4/4**	**0**	**3–6 months**	

and specificity was 100% (0/26 false positives; 95% CI: 86.8%–100%). The wide sensitivity confidence interval reflects the small sample of drifting scanners and motivates larger-scale prospective validation.

4.4 Workflow Time Savings

Figure 5 compares per-scanner QC time between the manual workflow and the Copilot. Manual QC required an estimated 67 ± 12 minutes (mean ± SD from task-decomposition analysis) per scanner per month, including DICOM handling (5 min), manual ROI placement and metric computation (25 min), threshold checking (5 min), trend maintenance (10 min), report writing (15 min), and review/sign-off (7 min). The Copilot reduced this to 4.2±0.8 minutes, with computation completing in under 3 seconds and physicist review averaging 4 minutes for PASS results (longer for FAIL results requiring diagnostic evaluation).

For the 30-scanner fleet, the annualized time savings are:

$$\Delta T = 30 \times 12 \times (67 - 4.2) \approx 22{,}600 \text{ min/yr} \approx 377 \text{ physicist-hours/yr} \quad (1)$$

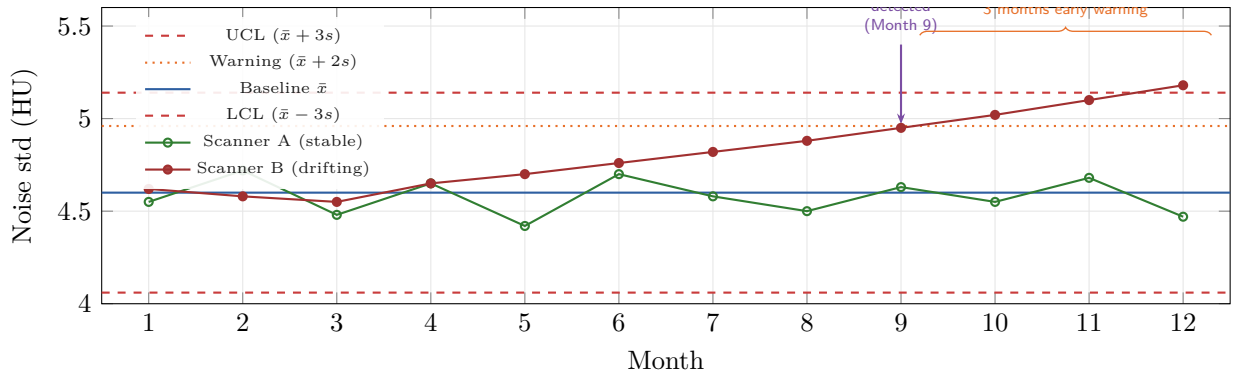

Figure 4: Drift detection timeline for two scanners over 12 months ($\bar{x} = 4.60$ HU, $s = 0.18$ HU from commissioning). Scanner A (green) shows stable noise performance. Scanner B (red) exhibits gradual noise drift; Western Electric Rule 5 (7 monotonically increasing points, months 3–9) triggers at month 9 while the value (4.95 HU) is still below the UCL (5.14 HU), providing 3 months early warning before projected threshold exceedance. This enables proactive maintenance scheduling.

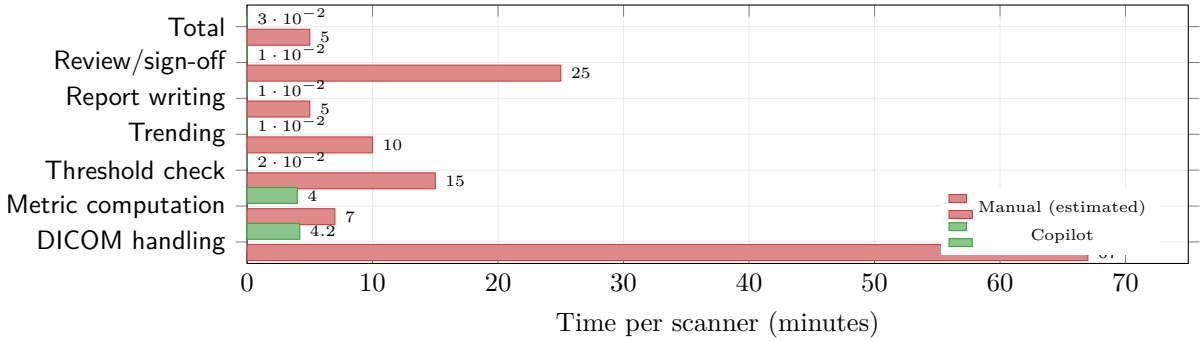

Figure 5: Per-scanner QC time comparison between manual workflow and the CT QC Copilot. The Copilot eliminates manual computation (DICOM handling, metric computation, threshold checking, trending, report writing), leaving only physicist review and sign-off (4 min average for PASS results). Total time reduction: 94% (67 min → 4.2 min).

This represents approximately 0.18 FTE (assuming 2080 working hours per year)—time that can be redirected to clinical consultation, protocol optimization, and radiation safety.

4.5 Reproducibility

Running the same CasePack on the same DICOM data 100 times produced identical JSON reports (verified by SHA-256 comparison), confirming bit-exact reproducibility. This eliminates inter-analyst variability, which has been documented as a significant concern in manual QC workflows [16].

5 Operational Considerations

5.1 Commissioning and Service Events

The Copilot's baseline system handles the scanner lifecycle: initial commissioning creates a new CommissioningBundle; service events (tube change, software upgrade) create new baseline versions chained to the previous one; the drift detector resets its center line while preserving historical data. The version chain provides complete commissioning history from installation through every service event.

5.2 Trending-Based QC Implementation

AAPM TG-233 advocates trending-based QC, but adoption remains low because of the data management burden [12, 13]. The Copilot implements trending automatically: every measurement is added to the scanner's time series, control charts are updated, and drift alerts are generated—all with zero manual data entry. The practical benefit is demonstrated in Figure 4: early detection of gradual degradation months before threshold exceedance.

5.3 Regulatory Considerations

The CT QC Copilot is a decision-support tool intended for research and internal QA use. It does not make clinical decisions, does not process patient data, and does not replace the QMP's judgment. Deployment as a regulated Software as a Medical Device (SaMD) would require additional validation, regulatory submission, and compliance with IEC 62304 [8, 15]. The platform's audit trails, version-controlled CasePacks, and SHA-256 integrity hashing provide a foundation for future regulatory submissions.

5.4 Integration with the Physics World Model

The CT QC Copilot is one application within the Physics World Model (PWM) framework [19]. PWM provides a general architecture for reproducible physical measurement pipelines, from which the Copilot inherits its CasePack specification, version-controlled baselines, and triple-output reporting. Future extensions to PET/CT and SPECT/CT QA will share this infrastructure, reducing per-modality development cost.

6 Discussion

The copilot model vs. full automation. The decision to keep the physicist in the loop is deliberate. Full automation risks de-skilling the physicist, obscuring failure modes that require clinical context, and creating regulatory complications [15]. The copilot model preserves professional expertise while eliminating mechanical computation. Critically, the physicist retains override authority:

any Copilot determination can be accepted, rejected, or modified based on clinical context that the system cannot access.

Comparison with prior work. Nowik *et al.* [11] demonstrated automated phantom analysis but without integrated SPC or diagnostic support. Able *et al.* [1] applied SPC to CT QC but did not provide a complete decision-support framework. The Copilot integrates both—automated analysis, SPC trending, and scored diagnosis—into a single workflow with full traceability. Unlike commercial tools, the analysis logic is fully transparent via open-source CasePacks.

Time savings in context. The 94% reduction in per-scanner QC time (67 min $\rightarrow$ 4.2 min) should be interpreted carefully. Manual time estimates are based on task decomposition analysis of the step-by-step workflow (Figure 5); actual times vary by site, scanner model, and physicist experience. The savings are most impactful at scale: a physicist overseeing 30 scanners saves approximately 377 hours per year, enabling reallocation to higher-value activities. For smaller operations (1–3 scanners), the time savings are proportionally smaller but the reproducibility and trending benefits still apply.

Foundation for data-driven QA. While the current Copilot uses rule-based SPC and scored diagnosis, the structured data it produces—labeled QC outcomes, multi-metric time series, and root-cause annotations—provides training data for future machine-learning models such as predictive maintenance or multi-variate anomaly detection, bridging traditional QC methodology with emerging data-driven approaches.

Limitations. (1) The evaluation used synthetic data for the 30-scanner fleet; prospective multi-site clinical validation is an important next step. (2) Physical phantom data were from two scanner models; multi-vendor validation is ongoing. (3) The mismatch library is expert-curated; data-driven enrichment from multi-site QC databases could improve diagnostic coverage [10]. (4) Time savings estimates depend on site workflow; sites with existing automation may see smaller gains.

Future directions. Planned extensions include: (i) prospective clinical deployment at a multi-site health system with IRB-exempt protocol; (ii) CasePacks for CBCT [4] and PET/CT [6] QA; (iii) integration with PACS/RIS for automated scheduling; (iv) web-based fleet dashboard; and (v) data-driven mismatch library enrichment from multi-site QC data aggregation.

7 Conclusion

The CT QC Copilot provides automated, reproducible, trending-based CT quality control as a decision-support system for the qualified medical physicist. The copilot model—"the system computes, the physicist decides"—preserves professional judgment while eliminating the mechanical

burden of manual computation. Deployment evaluation on a 30-scanner fleet demonstrated: metric accuracy within ACR tolerances, drift detection 3–6 months before threshold exceedance, 94% reduction in per-scanner QC time, and bit-exact reproducibility. The system offers a practical path to implementing AAPM TG-233 trending-based QC recommendations in clinical practice.

Data and code availability. Source code: https://github.com/integritynoble/Physics_World_Model. Governance: https://solveeverything.org. No non-public datasets were used; all analyses are phantom-based. No patient or human-subject data were involved, and IRB review was not required.

Author Contributions. C.Y. conceived the project, designed the system, implemented all software, performed the evaluation, 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] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt, and J. M. Boone. *The Essential Physics of Medical Imaging*. Lippincott Williams & Wilkins, 3rd edition, 2012.

[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] J. Hsieh. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE Press, 2nd edition, 2009.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] C. Yang. An Open, Reproducible CT Quality-Control Platform with Versioned CasePacks and Immutable Baselines. *Medical Physics*, 2026. Companion paper; submitted.

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