This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

S1. Confidence-Ranked Scan Reference Time Resolver

The scan reference time t_s was selected from a ranked list of DICOM-derived candidates (Table 3). When `DecayCorrection = ADMIN`, $t_s = t_i$ (HIGH). For `DecayCorrection = START`, candidates were assembled in the order of Table 3.

Table 3. Confidence-ranked candidates for scan reference time t_s.

Confidence	Source
HIGH	Enhanced PET `DecayCorrectionDateTime` (0018,9701)
HIGH	Vendor private tag: Siemens (0071,1022) or GE (0009,100D)
MEDIUM	Earliest `AcquisitionDateTime`/`AcquisitionTime`
MEDIUM	First-file `AcquisitionDateTime`/`AcquisitionTime`
LOW	`SeriesDate` + `SeriesTime`
VERY LOW	`SeriesDate`/`AcquisitionDate` + `StudyTime`

Vendor private tags were rated HIGH because they are designed to record this datetime and can recover t_s when `SeriesTime` is overwritten [7], but were not accepted blindly: a known Siemens failure mode stores the correct time of day with the wrong date, silently adding 24 h to the uptake interval [7]. The pipeline detects this by cross-checking each vendor datetime against the earliest acquisition datetime and repairing the date when a nearby DICOM date matches better; confidence is then downgraded to MEDIUM. Uptake time was $t_s - t_i$; negative values from midnight-crossing timestamps were corrected by adding 24 h when within 12 h, else flagged INVALID. Resolved t_s, t_i, confidence, source tag, repair status, and timing cross-check deltas were saved per scan for traceability.

S2. DICOM Fields and SUV Conversion

PET and radiopharmaceutical metadata were parsed from the first PET slice: units (0054,1001), decay-correction mode (0054,1102), corrected-image flags (0028,0051), rescale slope/intercept (0028,1053/0028,1052), SUV type (0054,1006), patient weight (0010,1030), and the radiopharmaceutical sequence (dose, half-life, injection time). Missing weights were filled from a study spreadsheet (source logged); weight came from DICOM in 115/130 scans. Units = SUV were treated as pre-converted; GML with SUV Type = BW as $\mathrm{SUV_{bw}}$. BQML series were converted only when `CorrectedImage` included `DECY` and all quantitative fields were present. Negative PET voxels were clipped to zero before resampling.

S3. Full QC Verdict Criteria

Image QC. FAIL — body Dice < 0.60 or edge-contact > 0.15. **WARN** — Dice $[0.60, 0.80)$; edge-contact $(0.05, 0.15]$; implausible weight ($< 30/> 250$ kg), height ($< 1.2/> 2.2$ m), or BMI ($< 12/> 60$). **PASS** otherwise.

SUV QC. FAIL — BQML not safely convertible; timing `MISSING`/`INVALID`; or `SeriesTime`-vs-`AcquisitionTime` > 30 min. **WARN** — liver $\mathrm{SUV}_{\mathrm{mean}} \notin [1.5, 2.5]$; `SeriesTime`-vs-`AcquisitionTime` 5–30 min; or a repaired private-tag date. **PASS** otherwise.
The combined verdict was FAIL if either component failed, WARN if neither failed but at least one warned, and PASS otherwise.

S4. Liver Reference VOI

A liver mask was generated from CT with TotalSegmentator. Within an eroded mask, a 30-mm spherical VOI was placed at the location of lowest within-sphere standard deviation; liver $\mathrm{SUV}_{\mathrm{mean}}$, SD, and CoV were recorded. The 30-mm geometry follows EARL/PERCIST recommendations [3, 15]; 3-cm spherical liver $\mathrm{SUV}_{\mathrm{mean}}$ is the most reproducible liver uptake metric [17]. The EANM v3.0 expected range is 1.3–3.0 [3]; we used the tighter 1.5–2.5 window based on Boktor et al. [4] (mean 2.17–2.29, SD ≈ 0.44). Liver QC was skipped for unconverted BQML scans.

S5. Matched-Random Control and Validation

To separate a QC-specific effect from sample composition, control subsets were matched to the centre $\times$ vendor composition of the PASS-only subset ($n = 78$ for the stability control; $n = 74$ for the prediction control, which was restricted to patients with a pCR endpoint) but drawn without regard to QC tier; per-cell counts were fixed to PASS-only and patients sampled at random within cells. We drew 100 subsets each for the stability and prediction controls, recomputing the full analysis on each. $P(\text{matched} \geq \text{PASS})$ is the fraction of draws meeting or exceeding the PASS-only statistic. $\mathrm{SUV}_{\mathrm{bw}}$ harmonization was validated against the ACCURATE PET research interface at the liver region; voxel-wise comparison was infeasible due to inconsistent headers and orientations across pipelines, and the 3-cm spherical liver $\mathrm{SUV}_{\mathrm{mean}}$ is the appropriate low-sensitivity agreement target.

S6. Image-QC Body Masks

Body masks were compared on the CT grid (PET linearly resampled to CT). The CT mask used HU > -400, morphological closing, hole filling, and largest-connected-component selection. The PET mask used SUV > 0.2 (SUV-compatible) or $> 10^{-3}$ Bq/mL (unconverted BQML) to avoid unit-driven Dice failures.