Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Target-Lesion Measurements Show High Aggregate Agreement but Limited Representation of Non-Target Response in Metastatic Melanoma

Haoran Sun
Beihang University
Kangqi Tian
Beihang University
Miaowen Jiang
Capital Medical University
Shiqiang Zheng
zhengshiqiang@buaa.edu.cn

Beihang University

Research Article

Keywords: Melanoma, Neoplasm metastasis, Computed tomography, Treatment response, Tumour burden

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11343363/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Target-Lesion Measurements Show High Aggregate Agreement but Limited Representation of Non-Target Response in Metastatic Melanoma

Haoran Sun[1], Kangqi Tian[1], Miaowen Jiang[2*], and Shiqiang Zheng[1*]

[1] School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China.
[2] Beijing Institute of Brain Diseases, Capital Medical University, Beijing, 100053, China.

ORCID: Haoran Sun, https://orcid.org/0009-0000-5819-4192; Kangqi Tian, https://orcid.org/0000-0002-1619-9777; Shiqiang Zheng, https://orcid.org/0000-0002-2842-9699.

Corresponding authors

Shiqiang Zheng: zhengshiqiang@buaa.edu.cn

Miaowen Jiang: jiangmiaowen@ccmu.edu.cn

Postal addresses for correspondence:
Shiqiang Zheng: School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China.
Miaowen Jiang: Beijing Institute of Brain Diseases, Capital Medical University, Beijing, 100053, China.

**Corresponding authors.*

Abstract

Background. To determine how well radiologist-designated target lesions represent whole-body tracked-lesion response in metastatic melanoma and whether limited sampling underrepresents intrapatient mixed response.

Methods. We used the Longitudinal-CT dataset, comprising paired whole-body CT scans from 300 patients with metastatic melanoma and 7,182 manually segmented, longitudinally linked lesions; 264 patients entered the primary analysis and 250 the common strategy cohort. Target, non-target tracked-lesion, and all-lesion changes were evaluated using axial diameter and volume. Response-direction discordance and mixed response were assessed at ±20%, ±30%, and ±50% thresholds, with a 2-mm safeguard for diameter. Sensitivity analyses examined effective target count, constrained randomized lesion-selection benchmarks, paired patient-level bootstrap estimates, and conditional permutation analyses of structural subset coupling.

Results. Target–all correlations were high (ρ=0.873 for diameter and 0.894 for volume); direct target–non-target correlations were 0.659 (n=228) and 0.678 (n=229), with 29.4–34.6% and 21.4–28.0%

direction discordance. Among 180 patients with at least two eligible diameter targets, mixed response remained unrepresented in 43/60 (71.7%), 35/46 (76.1%), 26/32 (81.3%). The mean diameter correlation of count-matched random selection was 0.854 (Monte Carlo 95% range, 0.811–0.889). Organ-balanced selection reduced diameter mean absolute error by 2.07 percentage points (95% CI, -4.05 to -0.20); threshold-specific discordance and mixed-response capture varied across strategies. Conditional shuffling retained positive target–all correlations below the observed values.

Conclusions. Limited target sampling preserved aggregate ordering while incompletely representing complementary tracked lesions and mixed response. Information loss persisted with at least two targets and under matched randomized benchmarks. Whole-body lesion information may complement conventional target-lesion assessment.

Keywords

Melanoma; Neoplasm metastasis; Computed tomography; Treatment response; Tumour burden

Background

RECIST reduces a spatially complex disease state to a reproducible patient-level summary by measuring a limited number of lesions. RECIST 1.1 reduced the maximum number of target lesions to five overall and two per organ while retaining separate assessment of non-target and new lesions [1]. This operational compromise supports trials and routine care, but its quantitative target-lesion component is necessarily a sample rather than an exhaustive representation of visible disease [2,3].

Sampling may be consequential in metastatic melanoma, where lesions within one patient can respond differently across organs and over time. Immunotherapy has also produced atypical trajectories that motivated immune-related criteria and iRECIST [4–6]. Target selection and baseline ambiguity can alter patient-level response assessment [7,8]. Inter-reader variability is also well documented in RECIST-based evaluation [9–11]. Yet high correlation between a selected subset and total burden can coexist with substantial individual disagreement. Moreover, the all-lesion summary contains the target subset; its correlation with target change is therefore partly structural, particularly when targets cover a large share of baseline burden. Direct comparison with the unselected lesions is needed to isolate this coupling.

A second issue is measurement dimensionality. One-dimensional diameter and three-dimensional volume are not geometrically interchangeable [12–14]. Both measurements are also subject to technical and observer variability [15,16]. Directly comparing target diameter with all-lesion volume would therefore confound sampling with measurement-system effects.

The Longitudinal-CT dataset provides 7,182 manually segmented and longitudinally linked lesions in 300 patients with metastatic melanoma [17]. We audited radiologist-designated targets using target, non-target, and all-lesion diameter (TD, NTD, and AD) and volume (TV, NTV, and AV) change. We asked whether targets represented unselected lesions after accounting for subset coupling, how often response directions differed, how much lesion-level mixed response was underrepresented, and whether simple interpretable selection rules consistently narrowed these gaps.

Methods

Study design and cohort

We used the Longitudinal-CT dataset [46], which comprises patients with malignant melanoma who underwent systemic therapy at University Hospital Tübingen between 2015 and 2019. The dataset includes 300 patients with paired baseline and follow-up whole-body CT examinations, 4,079 baseline lesions, and 3,103 follow-up lesions with manual segmentation and longitudinal lesion linkage. The source study was approved by the institutional ethics committee of the University of Tübingen (project number 700/2024BO2), which permitted the use and publication of anonymized imaging data; the requirement for informed consent was waived because data processing and analysis were anonymized. We followed the STROBE reporting principles [18].

The primary cohort excluded patients with no designated target lesion, ambiguous longitudinal lesion linkage, or merge groups containing both target and non-target baseline lesions. These overlapping criteria excluded 13, 12, and 12 patients, respectively; 36 unique patients were excluded, leaving 264 patients for the primary analysis. The common cohort used for deterministic lesion-selection comparisons further excluded patients with merge topology and those exceeding the two-targets-per-organ constraint, resulting in 250 patients.

We characterized target-lesion selection according to the number of designated targets, compliance with the five-lesion and two-per-organ limits, overlap with the largest lesions, the number of represented anatomical categories, and baseline target-volume coverage. Lesion count, anatomical-category count, and target coverage were summarized to characterize the extent and distribution of lesion sampling. New lesions were summarized separately at the patient level and were not included in change scores whose denominators were defined from baseline lesions.

Lesion topology and measurements

Persisting lesions were measured at both time points, disappearing lesions received a follow-up value of zero, and new lesions were reported separately rather than entered into baseline-denominated change. Homogeneous merge groups were aggregated in the volume framework because volume is additive. Merge groups were excluded from diameter aggregation because summing several baseline diameters and comparing them with one merged diameter lacks geometric equivalence. Consequently, TD–TV and AD–AV were explanatory sensitivity analyses rather than strict measurement conversions.

The topology audit identified 2,506 persisting, 1,407 disappearing, 559 new, and 166 merging lesion relationships. A lesion entered the primary tracked set only when its baseline–follow-up relationship was explicit. Patients with unclear linking were excluded from the primary analysis, rather than allowing the denominator or lesion set to vary silently between examinations.

The primary diameter was the mask-derived axial longest diameter in millimetres. Foreground voxels were transformed through the NIfTI affine to physical coordinates; the maximum in-plane convex-hull distance across native slices was retained. Synthetic validation covered geometry, affine orientation, anisotropy, boundary cases, and separated components. This was not a blinded manual RECIST reread.

Mask volume equalled the foreground voxel count multiplied by the absolute determinant of the 3×3 spatial affine; all 670 NIfTI masks received explicit read and mapping results. Measurement-variability evidence supported the mixed-response safeguard [19–22]. The primary diameter multicomponent rule retained the largest 26-connected component at ≥95% volume share and otherwise excluded the label; 99%, all-excluded, and raw-label policies were sensitivity analyses. Primary volume analyses retained the

dataset CSV lesion volumes and topology aggregation. Recomputed mask volumes served as an independent reconciliation check; affected-patient exclusion was a separate sensitivity analysis.

Quality control covered mask validity, affine and dimensions, label/file identity, duplicate identifiers, volume reconciliation, synthetic edge cases, random lesions, outliers, and discordant cases with overlaid measurement lines.

Response summaries and statistical analysis

For patient *i*, aggregate change was

$$\Delta_i = \frac{\Sigma_j M_{ij,FU} - \Sigma_j M_{ij,BL}}{\Sigma_j M_{ij,BL}} \quad (1)$$

where *M* was axial diameter or volume and *j* indexed lesions in the target, non-target, or all-tracked set. Because targets are a subset of all lesions, each same-metric comparison obeys

$$\Delta_{all,i} = w_i \Delta_{target,i} + (1 - w_i)\Delta_{non-target,i} \quad (2)$$

where w_i denotes the proportion of baseline diameter or volume burden represented by target lesions. This decomposition separates the target and unselected tracked-lesion contributions to all-lesion change and was used to evaluate the structural coupling between target and all-lesion summaries. We therefore assessed lesion-sampling representativeness primarily through direct TD–NTD and TV–NTV comparisons. Throughout the analysis, non-target lesions refer to longitudinally tracked lesions that were not selected as targets. Direct target–non-target comparisons were performed in patients with a nonzero baseline burden in the unselected lesion set, with corresponding target–all comparisons calculated in the same patients. Target-volume coverage was stratified as <25%, 25–<50%, 50–<75%, and ≥75%.

Associations between target and comparison summaries were assessed using Spearman correlation, and absolute disagreement was summarized using mean absolute error, median absolute error, and IQR. Confidence intervals for Spearman correlation coefficients were obtained from 3,000 patient-level percentile bootstrap resamples; proportions were summarized with Wilson 95% confidence intervals. Response direction was classified as decrease, stable, or increase at ±20%, ±30%, and ±50% thresholds. Diameter–volume sensitivity analyses additionally used geometrically corresponding spherical thresholds.

Mixed response was defined as the coexistence of at least one decreasing and one increasing tracked lesion beyond the selected threshold within the same patient. For diameter-based analyses, each qualifying change was additionally required to be at least 2 mm in absolute magnitude; volume mixed-response analyses used the percentage thresholds without an absolute-size safeguard. We quantified target capture among patients with all-lesion mixed response and among those with mixed response in the unselected lesion set. Sensitivity analyses excluded patients with new lesions, excluded the lowest quartile of baseline non-target burden, and required a baseline axial longest diameter or equivalent-sphere diameter of at least 5 or 10 mm. We also compared radiologist-selected targets with largest-diameter, organ-balanced, and largest-volume selection strategies under the same lesion-number and per-organ constraints.

Strategy comparisons were performed in a common cohort of 250 patients using identical lesion eligibility, topology handling, measurement definitions, and outcome metrics. Performance was evaluated at both the cohort and patient levels using correlation, absolute error, direction discordance, and mixed-response capture. Patient-level comparisons additionally quantified the proportion of cases in which each alternative strategy reduced, preserved, or increased target–all-lesion error relative to radiologist selection.

Sensitivity analyses

Sensitivity analyses were conducted to examine the robustness of the main findings to effective target count, alternative lesion-selection strategies, and structural coupling between target and all-lesion summaries, and to further characterize lesion-level information not captured by the aggregate target response. These analyses included target-count stratification, constrained randomized selection benchmarks, paired bootstrap comparisons, conditional permutation analysis, and assessment of non-target lesion growth, new-lesion overlap, and response divergence across anatomical categories.

Mixed-response robustness was assessed across strata defined by effective target count (1, 2, 3, and 4–5) and was further evaluated after restricting the analysis to patients with at least two eligible targets, because a single target lesion cannot by definition exhibit mixed response. Effective target count was

determined from usable longitudinal diameter pairs in the diameter framework and from valid volume topology groups in the volume framework; each homogeneous merged volume group was treated as a single analysis unit. Within each target-count stratum, we quantified the proportion of patients with all-lesion mixed response in whom mixed response was also captured by the target-lesion set, with Wilson 95% confidence intervals.

Lesion-selection strategies were further evaluated in the common 250-patient cohort using constrained randomized benchmarks. Count-matched random selection preserved each patient's number of target lesions while enforcing the five-lesion overall limit and the two-per-anatomical-category constraint. Category-matched random selection additionally preserved the number of targets selected from each anatomical category. For each strategy, 1,000 cohort-wide realizations were generated, with one admissible target subset selected for each patient in every realization; the corresponding non-target set was defined as the complementary eligible lesion set. An additional 1,000 realizations were used to assess Monte Carlo stability.

Strategy performance was compared using target–all absolute discrepancy, response-direction discordance, mixed-response capture, and patient-level changes in absolute error. Uncertainty in paired strategy contrasts was estimated using 3,000 patient-level bootstrap resamples. For randomized strategies, each bootstrap replicate incorporated one complete randomized realization, thereby accounting for both between-patient sampling variation and variability arising from lesion selection. Patient-level absolute-error differences within 10^{-12} fractional-change units were considered numerical ties.

Structural coupling between target and all-lesion response was further assessed by conditional permutation analysis. For each patient, target-lesion change and baseline target-burden share were preserved, while non-target change was permuted within strata defined jointly by baseline target coverage (<25%, 25–<75%, and ≥75%) and the number of involved anatomical categories (one versus ≥2). When a resulting stratum contained fewer than 10 patients, the anatomical-category strata were pooled within the corresponding coverage group. For each of 3,000 permutations, all-lesion change was reconstructed from the permuted non-target change using the weighted target/non-target decomposition defined above,

generating a reference distribution of target–all correlation for both diameter and volume. Coverage-stratified and fully unstratified permutations were additionally performed to assess the sensitivity of the results to the permutation scheme.

Lesion-level response information beyond the aggregate target summary was examined by identifying patients whose target-lesion response was stable or decreasing while at least one non-target tracked lesion showed growth beyond the corresponding response threshold. For diameter-based analyses, qualifying growth was additionally required to be at least 2 mm in absolute change. We then assessed the overlap between this response pattern and the occurrence of new lesions, and separately identified patients with opposing aggregate response directions across anatomical categories. New lesions were analyzed as an independent imaging feature and were not incorporated into change measures defined from baseline lesions. To illustrate the lesion-level pattern underlying target–all discordance, paired baseline and follow-up CT images from the case presented in Figure 5b were included in the Supplementary Material with the original segmentation contours and physical scale bars.

Use of generative AI in manuscript preparation

OpenAI ChatGPT and Codex were used to assist with language editing, manuscript formatting, and drafting portions of the Python analysis code. All analysis code, statistical outputs, interpretations, and manuscript content were reviewed and verified by the authors, who take full responsibility for the final work.

Results

Cohort and technical validation

Of 300 source patients, 264 formed the primary cohort and 250 formed the common strategy cohort (Fig. 1). Median baseline lesion count was 8 (IQR, 3–17.25; range, 1–59), target count was 2 (IQR, 1–3; range, 1–5), involved-organ count was 2 (IQR, 1–3; range, 1–7), and target-volume coverage was 59.9% (IQR, 28.1–93.1%). No patient had more than five targets; two exceeded two targets in an organ. Targets exactly matched the largest k lesions in 116/264 (43.9%); median largest-k overlap was 66.7% (IQR, 50.0–100%) and median target-organ coverage was 90.0% (IQR, 50.0–100%) (Table S1). Fifty-three

patients had new lesions. All 7,182 labels were reconciled. Twenty-three multicomponent labels affected 15 patients; component-policy changes altered discordant counts by at most one.

Independent remeasurement of all 670 masks recovered all 7,182 lesion labels, with maximum differences below 10^{-8} mm for diameter and 10^{-6} mm³ for mask-derived volume relative to the reference measurements. All 23 multicomponent labels reproduced their component counts and largest-component measurements, and connectivity was confirmed in 3,110 audited labels. Eighteen labels in 11 patients showed >1% differences between CSV and mask-derived volumes; excluding the eight affected patients in the primary cohort retained target–non-target correlations of 0.651 for diameter and 0.676 for volume (Table S10).

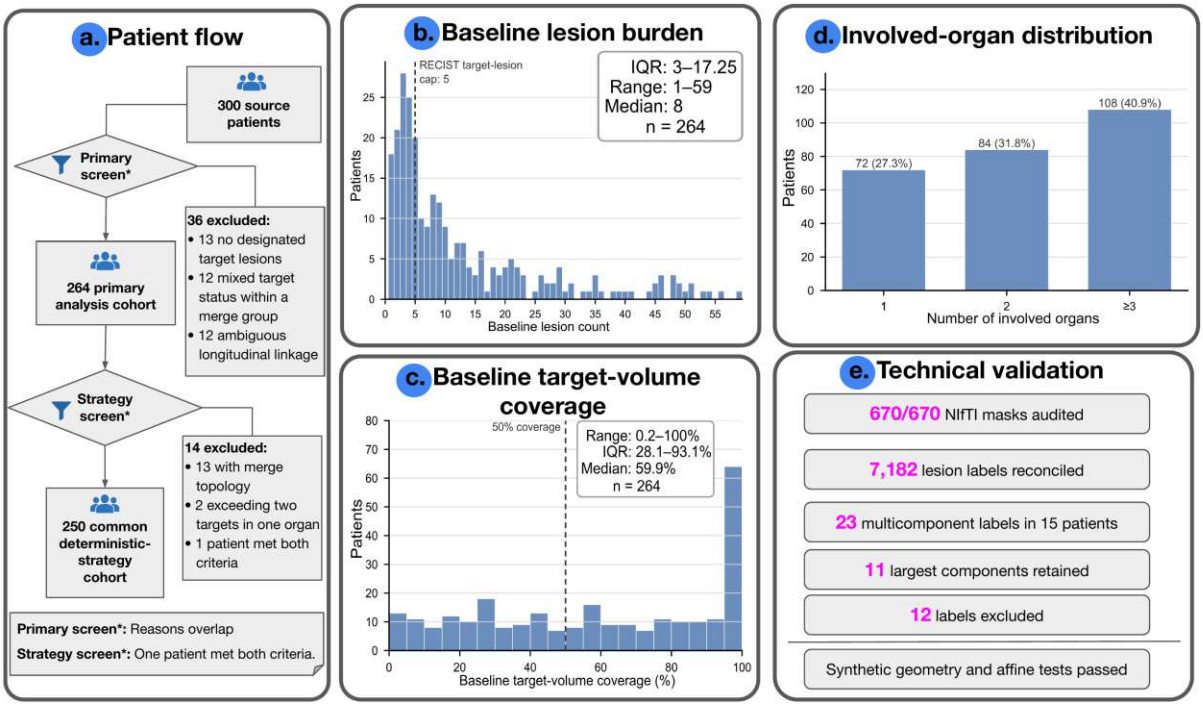

Figure 1. Cohort definition and technical validation. **a** Patient flow. **b** Baseline lesion-count distribution. **c** Target-volume coverage. **d** Involved organs. **e** Mask and multicomponent quality control. Exclusion counts overlap.

Baseline lesion counts were 1–5 in 112 patients, 6–10 in 53, 11–20 in 44, and >20 in 55; tracked disease involved one, two, or at least three organs in 72, 84, and 108 patients, respectively.

Overall agreement and patient-level discordance

In the full cohort, TD and AD were strongly associated (Spearman ρ=0.873; mean absolute error, 17.1 percentage points), as were TV and AV (ρ=0.894; mean absolute error, 54.8 percentage points) (Fig. 2; Table 1). However, the weighted target/non-target identities held to maximum absolute residuals of 4.44×10^{-16} for diameter and 7.11×10^{-15} for volume. Thirty-six patients had no eligible non-target diameter and 35 had no eligible non-target volume; for these patients, target and all-lesion summaries were identical by construction.

After restriction to patients with unselected tracked lesions, target–all correlations fell to 0.845 for TD–AD (n=228) and 0.875 for TV–AV (n=229). Direct TD–NTD correlation was 0.659 (95% CI, 0.565–0.741) and TV–NTV correlation was 0.678 (95% CI, 0.586–0.759). Mean absolute errors were 32.1 and 135.7 percentage points, whereas median errors were 22.8 (IQR, 10.4–41.7) and 23.0 (IQR, 7.7–84.4) percentage points, respectively. Direction discordance was 29.4–34.6% for TD–NTD and 21.4–28.0% for TV–NTV.

Coverage stratification supported the expected coupling mechanism. Across volume-coverage strata, TV–AV correlation rose from 0.790 below 25% coverage to 0.994 at ≥75%, whereas TV–NTV correlation remained 0.654–0.694. In the 19 diameter-eligible patients at ≥75% coverage, TD–AD correlation was 0.925 but TD–NTD correlation was 0.375. These descriptive strata were not used to define a high-risk subgroup.

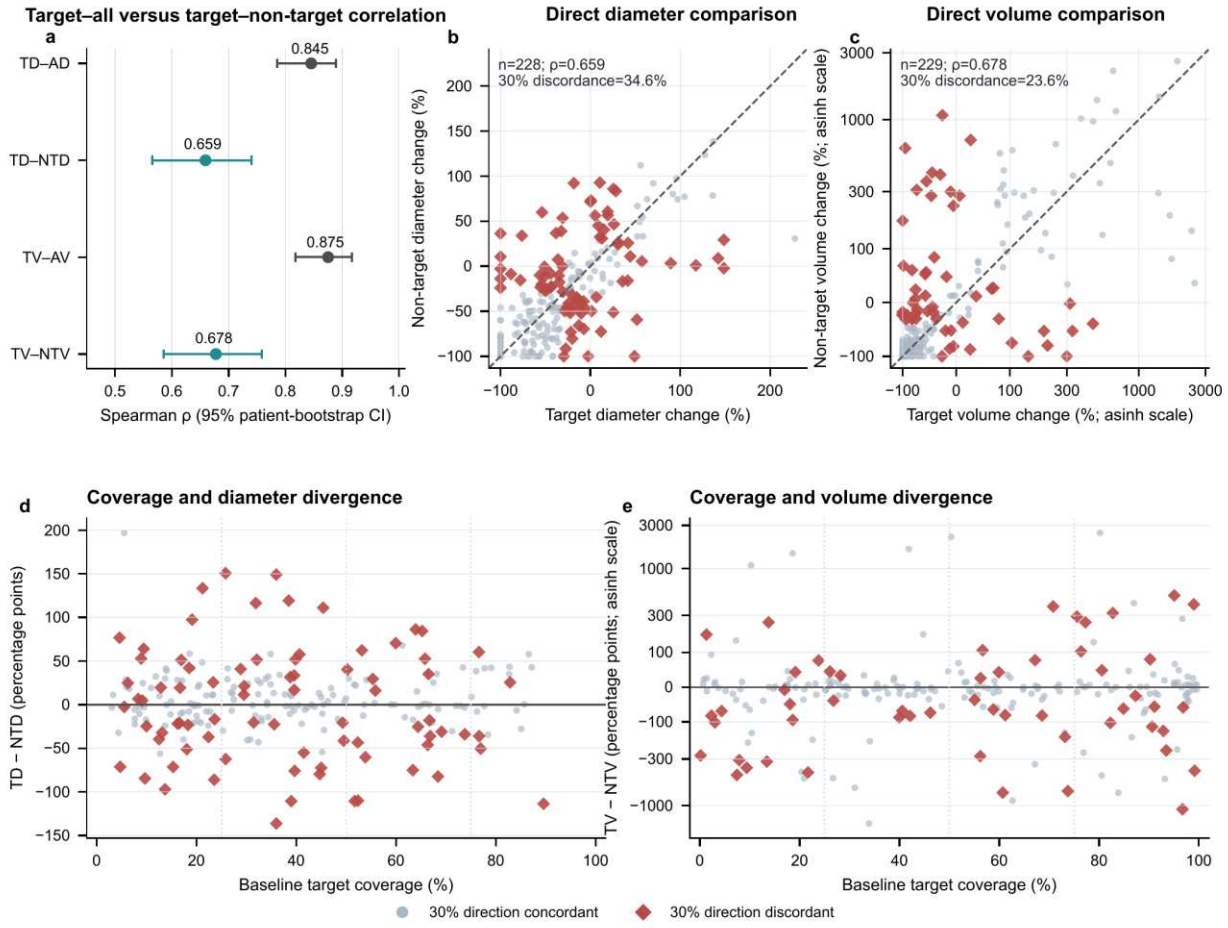

Figure 2. Direct target–non-target response comparison after accounting for structural subset coupling. **a** Spearman correlations for target–all and direct target–non-target comparisons; error bars indicate patient-bootstrap 95% confidence intervals. **b,c** Direct TD–NTD and TV–NTV comparisons; red diamonds denote 30% direction discordance. **d,e** Baseline target coverage versus target–non-target response divergence for diameter and volume. Volume axes use an inverse-hyperbolic-sine scale to retain all finite observations. Abbreviations: TD, target-lesion diameter change; AD, all-lesion diameter change; NTD, non-target tracked-lesion diameter change; TV, target-lesion volume change; AV, all-lesion volume change; NTV, non-target tracked-lesion volume change.

Excluding the 53 patients with new lesions retained 25.8–33.1% TD–NTD and 22.3–30.2% TV–NTV discordance. Excluding the lowest non-target-burden quartile or requiring baseline lesion/equivalent diameter of at least 5 or 10 mm likewise did not remove discordance (Tables S2–S3).

Conditional permutations retained positive target–all correlations. In the non-target eligible subsets, observed diameter correlation was 0.845, compared with a null median of 0.463 (95% permutation range, 0.368–0.561); corresponding volume values were 0.875 and 0.582 (0.498–0.659). Coverage-only and unstratified diagnostics gave similar distributions (Table S8; Fig. S2).

Mixed-response underrepresentation

With the 2-mm safeguard, all-lesion mixed response was present in 73, 58, and 40 patients across the three thresholds (Fig. 3; Table 2). Targets also showed mixed response in only 17, 11, and 6 patients; underrepresentation therefore affected 56/73 (76.7%), 47/58 (81.0%), and 34/40 (85.0%) patients. In 24, 20, and 14 patients, the only qualifying growing lesion was non-target. These values describe tracked-lesion heterogeneity, not missed RECIST progression.

Among patients whose non-target diameter lesions themselves showed mixed response, targets captured mixed response in 7/55 (12.7%), 4/45 (8.9%), and 2/34 (5.9%) patients across thresholds. Corresponding target capture among patients with non-target volume mixed response was 17/84 (20.2%), 16/78 (20.5%), and 8/61 (13.1%).

Eighty-four patients had one usable diameter target, and 180 had at least two. At the three thresholds, the single-target stratum contributed 13, 12 and 8 logically unavoidable unrepresented mixed responses. Among patients with at least two effective targets, underrepresentation remained 43/60 (71.7%; 95% CI, 59.2–81.5), 35/46 (76.1%; 95% CI, 62.1–86.1), 26/32 (81.3%; 95% CI, 64.7–91.1). Effective diameter target counts differed from designated counts in four patients; volume topology target counts were unchanged in this cohort. Results by effective count and framework are reported in Table S5 and Fig. S1 (Additional file 1).

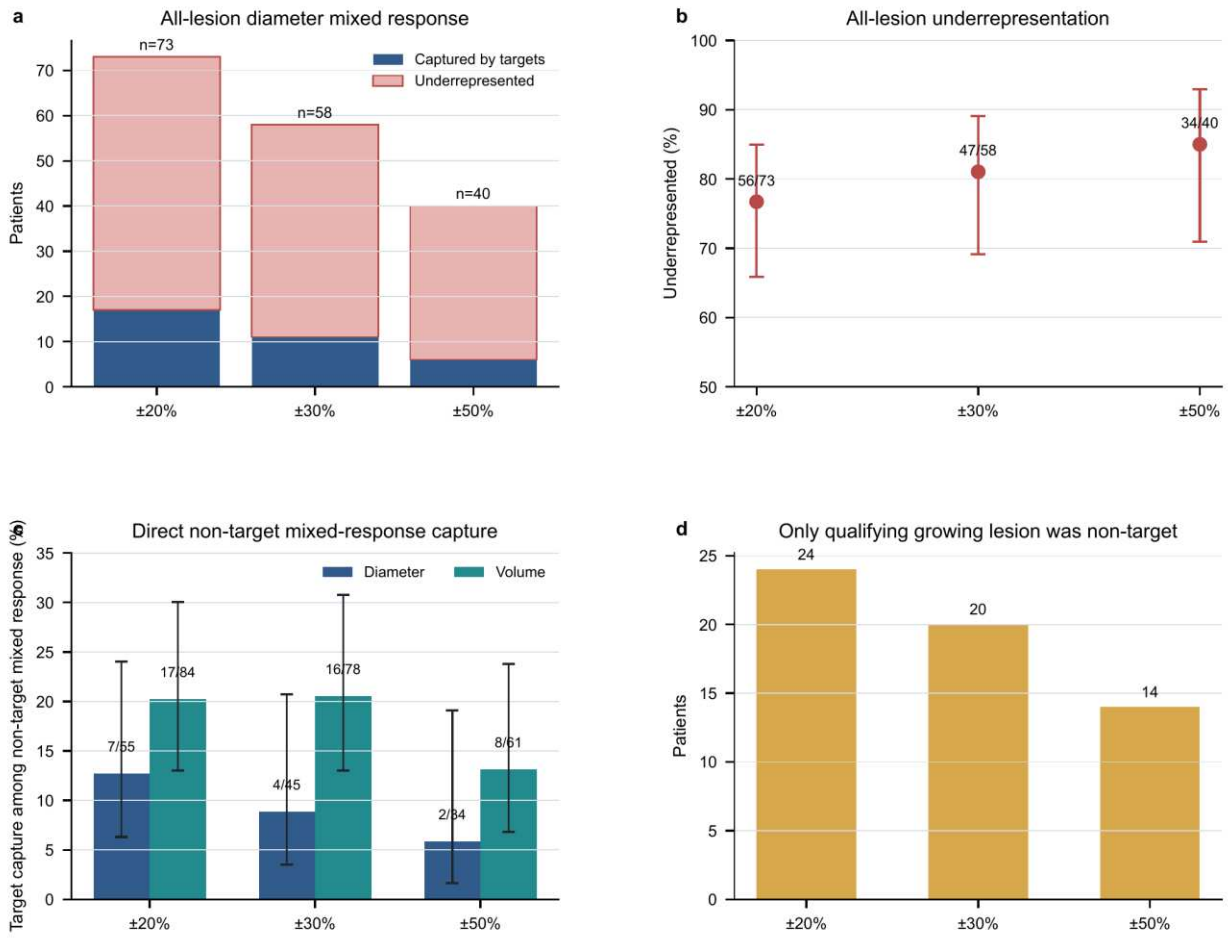

Figure 3. Mixed-response information loss. **a** All-lesion diameter mixed responses captured or underrepresented by targets. **b** Underrepresentation proportions with Wilson 95% confidence intervals. **c** Direct target capture among patients with mixed response in non-target tracked lesions, separately for diameter and volume; fractions show captured/eligible mixed-response patients and error bars are Wilson 95% confidence intervals. **d** Patients whose only qualifying growing diameter lesion was non-target. Diameter analyses required $|\Delta D| \geq 2$ mm. Here, non-target denotes eligible tracked lesions not designated as targets, not qualitative RECIST non-target assessment.

Measurement systems, cases, and selection strategies

TD–TV and AD–AV remained highly ranked (ρ=0.957 and 0.929), but identical numeric thresholds produced substantial category differences (Fig. 4). Geometrically corresponding thresholds reduced, but

did not eliminate, discordance. These analyses show non-interchangeability, not superiority of either metric [23–25].

At identical numeric thresholds, TD–TV discordance increased from 17.4% at 20% to 37.1% at 50%, and AD–AV discordance increased from 17.4% to 36.0%. With spherical geometric pairing, discordance was 14.0–15.5% for TD–TV and 14.8–21.2% for AD–AV. The RECIST-like asymmetric sensitivity yielded 45/264 (17.0%) and 47/264 (17.8%), respectively.

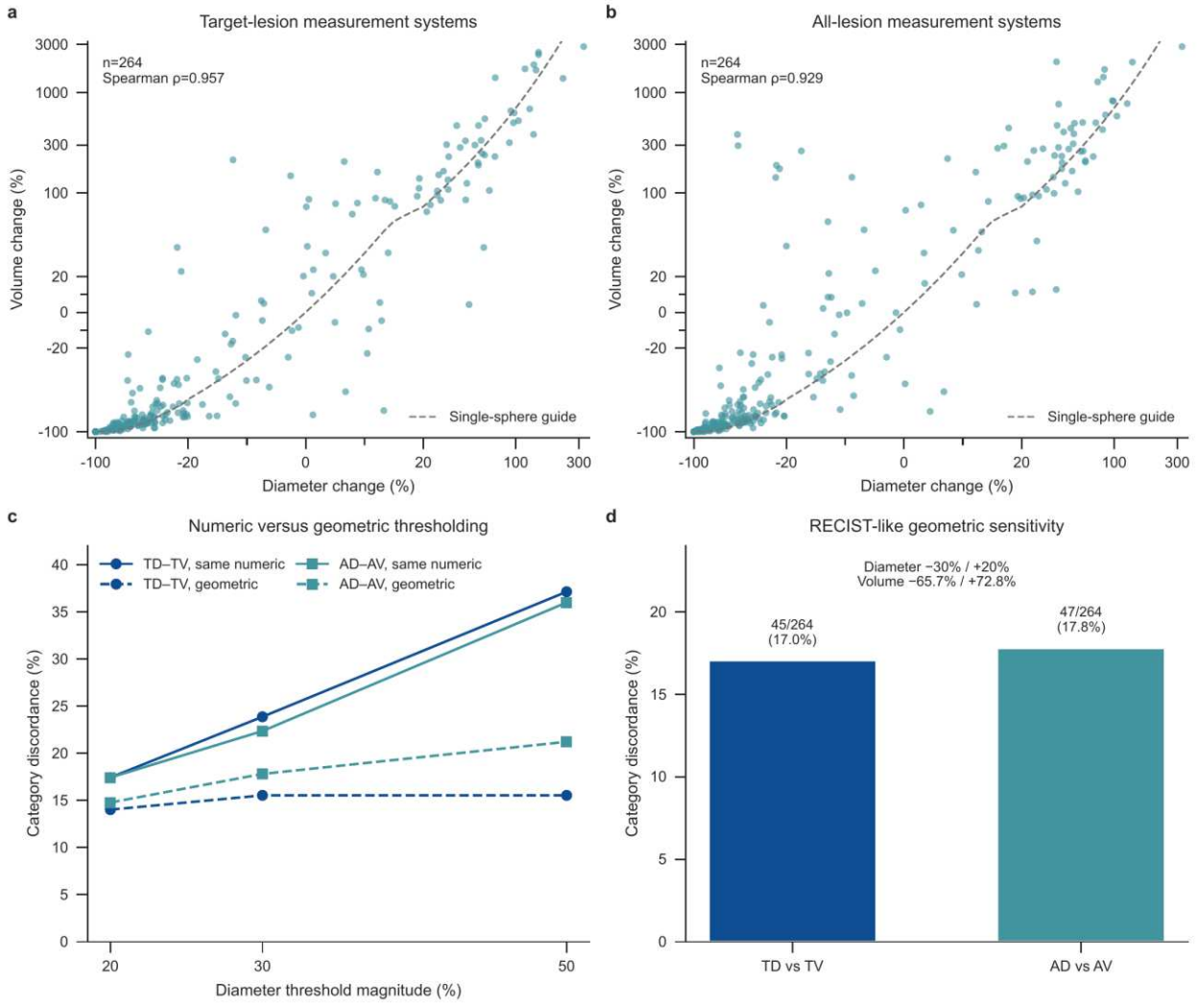

Figure 4. Diameter–volume sensitivity. **a,b** Target-set and all-lesion-set comparisons. **c** Identical numeric versus geometric thresholding. **d** RECIST-like asymmetric geometric sensitivity.

Four representative cases illustrate target–all discordance across different lesion counts (Fig. 5).

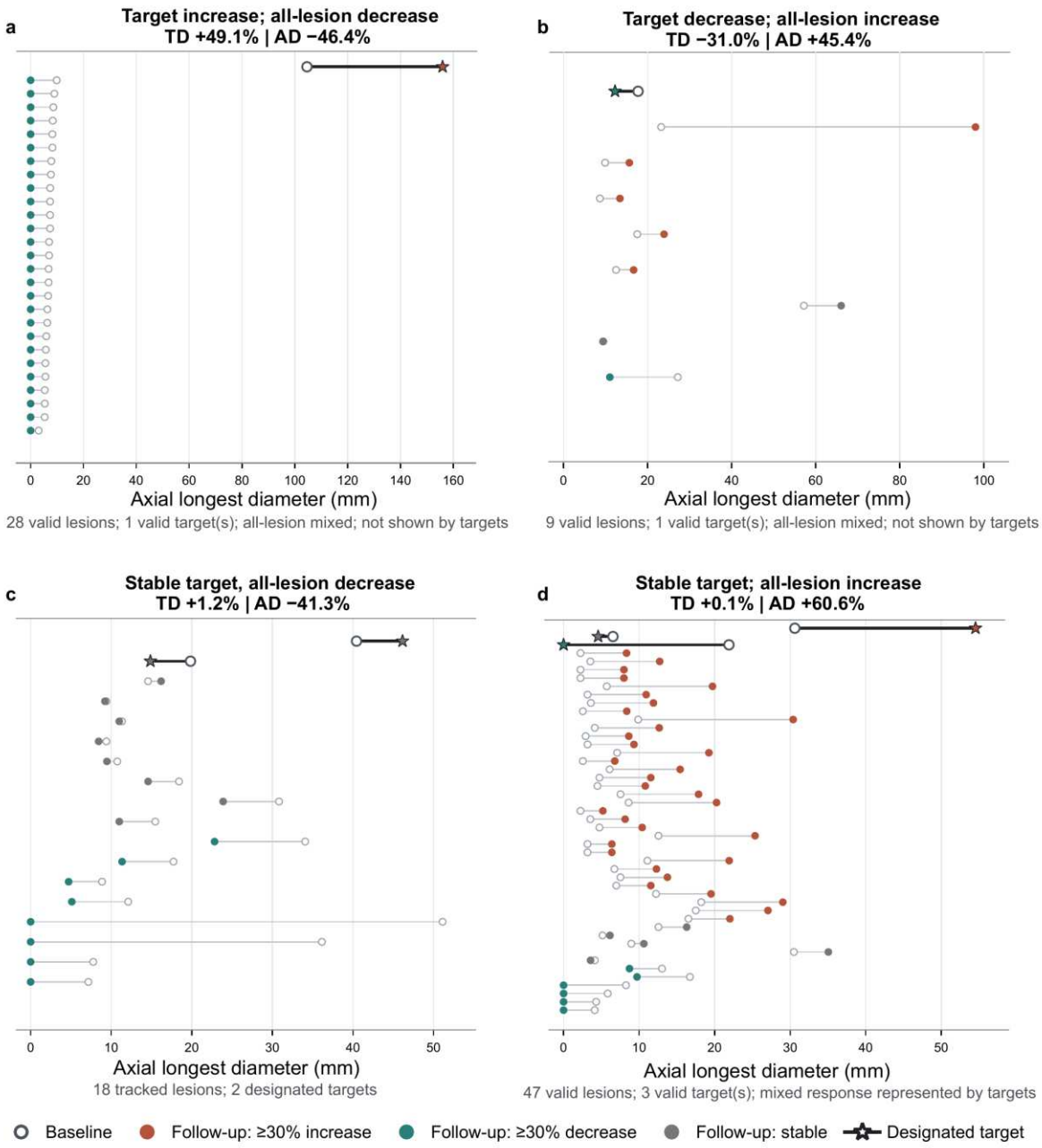

Figure 5. Lesion-resolved target–all discordance. Segments connect baseline and follow-up diameters; stars identify targets and colours indicate 30% relative categories. Panel c: TD +1.2%, AD −41.3%, 18 tracked lesions and two targets. Paired CT for panel b appears in Fig. S4. TD, target diameter change; AD, all-lesion diameter change.

The examples included both target increase with all-lesion decrease and the converse, across patients with few and many lesions (Fig. 5). Example TD/AD pairs included +49.1%/-46.4%, -31.0%/+45.4%, +1.2%/-41.3%, and +0.1%/+60.6%. They explain the quantitative mechanism but do not support outcome inference.

In 250 common patients, largest-diameter and organ-balanced rules improved average correlation and mean absolute error but not every direction-discordance or mixed-response endpoint (Fig. 6; Table 3). No deterministic rule was uniformly superior across thresholds.

For largest-diameter selection, diameter absolute error improved in 34.4% of patients, was unchanged in 43.2%, and worsened in 22.4%; the corresponding volume proportions were 38.8%, 44.4%, and 16.8%. Organ-balanced selection achieved the lowest average diameter and volume errors, yet its diameter discordance ranged from 15.2% to 19.2% and was highest at the 50% threshold. Thus, average gains did not identify one consistently better rule.

Using the 10^{-12} fractional-change numerical-tie tolerance, the corresponding largest-diameter diameter-error proportions were 34.4% improved, 44.4% tied and 21.2% worsened; three round-off-scale differences account for the change from the original exact-comparison proportions (Table S7).

Count-matched constrained random selection had mean diameter correlation 0.854 (Monte Carlo 95% range, 0.811–0.889) and mean volume correlation 0.812 (0.763–0.860). Its mean target–all errors were 18.9 and 85.7 percentage points. Exact organ matching yielded correlations 0.859 and 0.858. All 250 patients had feasible count and organ matching; the direct target–non-target comparison contained 215 patients (Table S6).

Compared with radiologist designation, organ balancing reduced mean absolute error by 2.07 diameter percentage points (95% CI, -4.05 to -0.20) and 13.12 volume percentage points (-23.92 to -4.86). Largest-diameter and largest-volume diameter contrasts had intervals spanning zero. Threshold-specific capture and discordance contrasts showed outcome-dependent trade-offs (Table S7; Fig. S3).

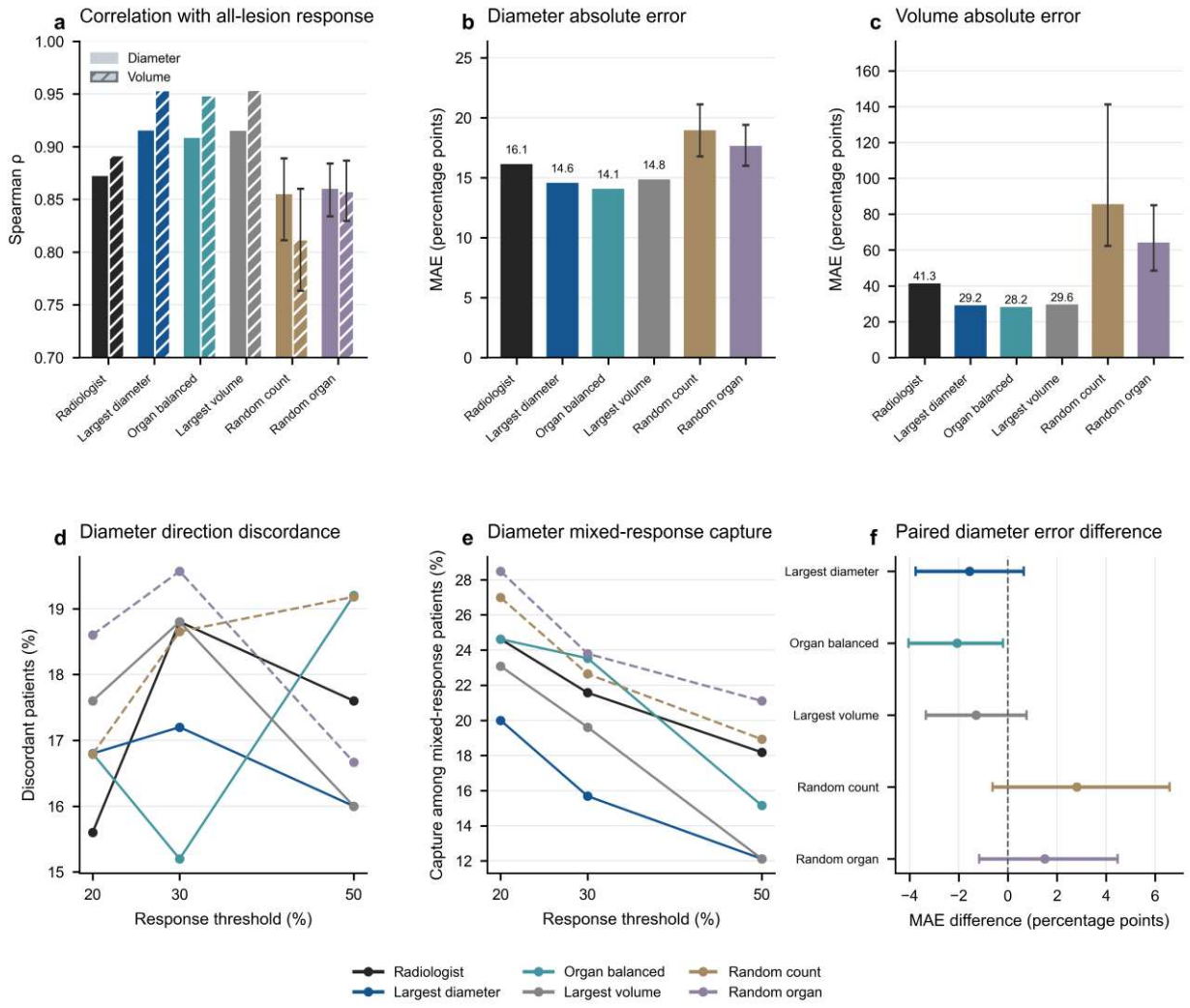

Figure 6. Matched lesion-selection benchmarks in 250 patients. a Target–all Spearman ρ; colours identify strategy, solid bars diameter, hatched bars volume. b,c Diameter and volume mean absolute error (MAE). d Diameter target–all discordance. e Diameter mixed-response capture; denominators are 65/51/33 at 20/30/50%. Random panels show means of 1,000 complete realizations; whiskers in a–c are Monte Carlo 2.5th–97.5th percentiles. f Diameter MAE differences versus radiologist with 3,000 patient-bootstrap 95% intervals; random intervals include complete-realization sampling variation. Negative differences indicate lower error. Random count matches k; random organ matches per-category counts.

At the 30% diameter threshold, 56/264 patients (21.2%; Wilson 95% CI, 16.7–26.5%) had stable or decreased target aggregates with at least one significantly growing non-target tracked lesion. Nineteen also had a new lesion, and 37 had no new lesion. Opposite aggregate directions across anatomical

categories occurred in 18/264 (6.8%) patients. New lesions overlapped all-lesion diameter mixed response in 21 patients, leaving 32 new-lesion patients without that mixed-response pattern (Table S9).

Discussion

The high target–all correlations cannot be interpreted as independent evidence of representativeness because targets are embedded in AD and AV. With median target-volume coverage of 59.9%, the verified weighted identity pulls the summaries together. Excluding patients without eligible unselected disease reduced target–all correlations; direct target–non-target correlations were 0.659–0.678, with 21–35% direction discordance. Targets therefore retained moderate co-response with unselected lesions, but less than unadjusted target–all coefficients suggested. Coverage-stratified divergence was consistent with the expected coupling mechanism.

This coupling analysis also reconciles favourable cohort ordering with patient-level error and 77–85% all-lesion mixed-response underrepresentation. Correlation measures ordering, whereas categorical concordance depends on the individual difference and its position relative to a threshold. RECIST reading studies have similarly shown that target selection and progression components create non-equivalent patient summaries [26,27].

The conditional shuffle adds evidence that positive target–all association can persist when within-stratum target–non-target co-response is broken. Observed correlations exceeded those distributions, supporting an additional empirical co-response signal. This diagnostic depends on exchangeability within coarse baseline strata and preserves some between-stratum structure. It therefore supports a qualitative distinction between subset coupling and observed lesion association; it cannot apportion correlation into causal percentages.

The largest information loss concerned lesion-level heterogeneity rather than wholesale reversal of aggregate trend. Mixed or dissociated response has been described using lesion-level and organ-specific response patterns [28–31]. With three prespecified thresholds and a 2-mm safeguard, targets captured only 5.9–12.7% of diameter and 13.1–20.5% of volume mixed responses arising within unselected tracked lesions. Interlesional divergence in melanoma has been associated with differential immune

infiltration, organ-specific biology, and intrapatient molecular heterogeneity [32–35]. Similar lesion-level response heterogeneity has been reported in other advanced cancers [36–38]. Without clinical outcomes, underrepresentation cannot be equated with missed progression, treatment failure, or altered survival.

Target count is a structural part of the mixed-response definition. A single target cannot display both qualifying growth and shrinkage. Removing that logical limitation reduced underrepresentation modestly, with 71.7–81.3% remaining among mixed-response patients who had at least two usable diameter targets. The results demonstrate residual information loss after count conditioning. Differences between target-count strata are descriptive because patients with different counts also differ in lesion burden and distribution.

The finding also clarifies why the target-lesion framework can remain useful. A finite subset can estimate the dominant aggregate trajectory while omitting minority lesion trajectories that cancel within the sum or arise outside sampled organs. The problem is therefore not simply that five lesions are "insufficient". Representativeness depends on baseline burden, organ distribution, target coverage, lesion-specific trajectories, and the particular endpoint being preserved. This narrower interpretation respects the operational purpose of RECIST while identifying information that its quantitative target component was not designed to retain.

Diameter and volume remained non-interchangeable because equal percentages do not represent equal geometry and summed diameters and total volume are not one physical object [12,14,39]. AV was a quantitative reference, not a biological gold standard; these results do not justify replacing diameter with volume.

Simple selection rules improved average error in some analyses [40,41], but none consistently reduced direction discordance and mixed-response omission. Higher average correlation alone cannot support replacing radiologist designation or training a complex selection model on this cohort.

The matched random benchmarks place radiologist designation within a fair sampling context. Original selection preserved aggregate response more closely on average than random count-matched selection,

especially for volume, while random subsets sometimes captured mixed response more often. Organ balancing improved average error in the paired analysis, but capture and threshold-specific discordance did not improve uniformly. These outcomes describe competing measurement objectives and provide no single rule with a consistent advantage across all examined endpoints. Selection simulations estimated the consequences of alternative target sets [8]; this audit measured actual responses of complementary lesions within exhaustive paired annotations. Whole-body volumetry studies established feasibility and reader variability [41]; the present analysis quantifies information retained by limited sampling within matched measurement frameworks.

Whole-body automation could complement target summaries with lesion-change distributions or discordance flags [42–45]; technical feasibility does not establish clinical substitution.

The additional lesion audit identifies complementary imaging features: growth among unselected tracked lesions, cross-category divergence, and new-lesion occurrence. Their partial overlap supports reporting these features alongside the target quantitative component. Their potential clinical relevance requires complete RECIST rereads and independent treatment or outcome information.

This study has limitations. Dataset target flags may not reproduce prospective RECIST measurability, lymph-node short-axis rules, qualitative non-target assessment, or clinical judgement. Diameters were mask-derived rather than independently reread, and merge handling differed by metric. Only baseline and one follow-up were available; treatment strata, external validation, and outcomes were unavailable. Relative volume change can be extreme with a small non-target denominator, although median errors, exclusion of the lowest burden quartile, 5/10-mm size restrictions, and exclusion of new-lesion patients retained frequent discordance. Neither unselected- nor all-lesion tracked response is a biological or clinical gold standard. Prospective multi-reader studies should combine complete RECIST, all-lesion quantification, treatment information, and outcomes. The added count, random and permutation experiments were exploratory; bootstrap intervals were unadjusted. Conditional shuffling may leave residual confounding within strata. Connectivity confirmation followed the original targeted audit rather

than exhaustive enumeration of every label. Dataset anatomical categories include broad groups, and the CT example was selected for illustration after analysis.

Strengths included exhaustive within-patient longitudinal lesion annotations, separate same-metric tests of sampling, explicit topology rules, affine-aware physical measurements, multicomponent sensitivity analyses, multiple prespecified response thresholds, source-data exports, and reproducible file hashes. These safeguards reduce the likelihood that the main findings arose from changing lesion denominators, cross-component diameter inflation, or a favourable post hoc threshold.

Conclusions

In conclusion, high target–all-lesion correlations were partly structurally induced. Direct target–non-target comparisons showed moderate correlation, frequent direction discordance, and poor capture of non-target mixed response. Whole-body quantitative summaries should be evaluated as a complement to, rather than a replacement for, conventional target-lesion assessment.

List of abbreviations

AD all-lesion diameter change

AV all-lesion volume change

NTD non-target tracked-lesion diameter change

NTV non-target tracked-lesion volume change

RECIST Response Evaluation Criteria in Solid Tumours

TD target-lesion diameter change

TV target-lesion volume change

Supplementary information

Additional file 1

File format: .docx

Title: Supplementary methods, tables and figures

Description: Supplementary Tables S1–S10 and Figures S1–S4 provide sensitivity analyses, target-count-conditioned analyses, randomized selection benchmarks, paired contrasts, coupling diagnostics, incremental imaging-information analyses, volume-outlier checks, and a paired CT example.

Additional file 2

File format: .zip

Title: Code and derived source data

Description: Analysis scripts, configuration files, tests, and derived CSV inputs used to reproduce the sensitivity analyses.

Declarations

Ethics approval and consent to participate

The source study was approved by the institutional ethics committee of the University of Tübingen (project number 700/2024BO2). The committee permitted the use and publication of anonymized imaging data and waived the requirement for informed consent because data processing and analysis were anonymized. The present retrospective analysis used only the publicly available, de-identified dataset and involved no participant contact or access to identifiable information.

Consent for publication

Not applicable. No identifiable individual participant information is presented.

Availability of data and materials

The source imaging data analyzed during this study are available as Longitudinal-CT v2 at the University of Tübingen FDAT Data Repository [46] and are described by Gatidis et al. [17]. Derived figure data and analysis code are available at https://github.com/hsun77/Longitudinal_CT_Target_Lesion_Audit. Analysis scripts, configuration files, tests, and derived source data for the sensitivity analyses are provided in Additional file 2. Raw CT images and segmentation masks are obtained from the original repository under its dataset licence.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant 62225301), the China Postdoctoral Science Foundation (No. 2026M794667), the Postdoctoral Fellowship Program of the China Postdoctoral Science Foundation (GZB20260887), and in part by the Beijing Science and Technology Planning Project (No. Z25110100040000). The funders had no role in the study design, data collection, data analysis or interpretation, preparation of the manuscript, or the decision to submit the manuscript for publication.

Authors' contributions

HS: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft. KT: Methodology, Validation, Writing – review & editing. MJ: Supervision, Writing – review & editing. SZ: Supervision, Funding acquisition, Writing – review & editing. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Authors' information

Not applicable.

References

1. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228–47. https://doi.org/10.1016/j.ejca.2008.10.026.
2. van Persijn van Meerten EL, Gelderblom H, Bloem JL. RECIST revised: implications for the radiologist. A review article on the modified RECIST guideline. Eur Radiol. 2010;20(6):1456–67. https://doi.org/10.1007/s00330-009-1685-y.

3. Suzuki C, Jacobsson H, Hatschek T, Torkzad MR, Bodén K, Eriksson-Alm Y, et al. Radiologic measurements of tumor response to treatment: practical approaches and limitations. Radiographics. 2008;28(2):329–44. https://doi.org/10.1148/rg.282075068.

4. Wolchok JD, Hoos A, O'Day S, Weber JS, Hamid O, Lebbé C, et al. Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. Clin Cancer Res. 2009;15(23):7412–20. https://doi.org/10.1158/1078-0432.CCR-09-1624.

5. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol. 2017;18(3):e143–e152. https://doi.org/10.1016/S1470-2045(17)30074-8.

6. Ramon-Patino JL, Schmid S, Lau S, Seymour L, Gaudreau PO, Li JJN, et al. iRECIST and atypical patterns of response to immuno-oncology drugs. J Immunother Cancer. 2022;10(6). https://doi.org/10.1136/jitc-2022-004849.

7. Iannessi A, Beaumont H, Liu Y, Bertrand AS. RECIST 1.1 and lesion selection: How to deal with ambiguity at baseline? Insights Imaging. 2021;12(1):36. https://doi.org/10.1186/s13244-021-00976-w.

8. Tareco Bucho TM, Tissier RLM, Groot Lipman KBW, Bodalal Z, Delli Pizzi A, Nguyen-Kim TDL, et al. How Does Target Lesion Selection Affect RECIST? A Computer Simulation Study. Invest Radiol. 2024;59(6):465–471. https://doi.org/10.1097/RLI.0000000000001045.

9. Yoon SH, Kim KW, Goo JM, Kim DW, Hahn S. Observer variability in RECIST-based tumour burden measurements: a meta-analysis. Eur J Cancer. 2016;53:5–15. https://doi.org/10.1016/j.ejca.2015.10.014.

10. Muenzel D, Engels HP, Bruegel M, Kehl V, Rummeny EJ, Metz S. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. Radiol Oncol. 2012;46(1):8–18. https://doi.org/10.2478/v10019-012-0009-z.

11. Iannessi A, Beaumont H, Ojango C, Bertrand AS, Liu Y. RECIST 1.1 assessments variability: a systematic pictorial review of blinded double reads. Insights Imaging. 2024;15(1):199. https://doi.org/10.1186/s13244-024-01774-w.
12. Sohaib SA, Turner B, Hanson JA, Farquharson M, Oliver RT, Reznek RH. CT assessment of tumour response to treatment: comparison of linear, cross-sectional and volumetric measures of tumour size. Br J Radiol. 2000;73(875):1178–84. https://doi.org/10.1259/bjr.73.875.11144795.
13. Prasad SR, Jhaveri KS, Saini S, Hahn PF, Halpern EF, Sumner JE. CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques initial observations. Radiology. 2002;225(2):416–9. https://doi.org/10.1148/radiol.2252011604.
14. Winter KS, Hofmann FO, Thierfelder KM, Holch JW, Hesse N, Baumann AB, et al. Towards volumetric thresholds in RECIST 1.1: Therapeutic response assessment in hepatic metastases. Eur Radiol. 2018;28(11):4839–4848. https://doi.org/10.1007/s00330-018-5424-0.
15. McNitt-Gray MF, Bidaut LM, Armato SG, Meyer CR, Gavrielides MA, Fenimore C, et al. Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error. Transl Oncol. 2009;2(4):216–22. https://doi.org/10.1593/tlo.09226.
16. Zimmermann M, Kuhl C, Engelke H, Bettermann G, Keil S. Volumetric measurements of target lesions: does it improve inter-reader variability for oncological response assessment according to RECIST 1.1 guidelines compared to standard unidimensional measurements? Pol J Radiol. 2021;86:e594–e600. https://doi.org/10.5114/pjr.2021.111048.
17. Gatidis S, Peisen F, Wagner A, Choudja POM, Othman A, Sanner A, et al. A longitudinal whole-body CT dataset with manually annotated tumor lesions. Sci Data. 2026. https://doi.org/10.1038/s41597-026-07466-y.
18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement:

guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344–9. https://doi.org/10.1016/j.jclinepi.2007.11.008.

19. Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. J Clin Oncol. 2011;29(23):3114–9. https://doi.org/10.1200/JCO.2010.33.7071.

20. Chung MS, Choi YJ, Lee SS, Lee JH, Lee YS, Rho JL, et al. Interobserver reproducibility of maximal axial diameter and tumour volume measurements from CT of patients with head and neck squamous cell carcinoma. Clin Radiol. 2017;72(11):993.e7–993.e13. https://doi.org/10.1016/j.crad.2017.06.006.

21. Gavrielides MA, Kinnard LM, Myers KJ, Petrick N. Noncalcified lung nodules: volumetric assessment with thoracic CT. Radiology. 2009;251(1):26–37. https://doi.org/10.1148/radiol.2511071897.

22. Guha S, Ibrahim A, Geng P, Wu Q, Chou Y, Akin O, et al. Variability of HCC Tumor Diameter and Density Measurements on Dynamic Contrast-Enhanced Computed Tomography. Tomography (Ann Arbor, Mich.). 2025;11(3). https://doi.org/10.3390/tomography11030036.

23. Gong AJ, Ruchalski K, Kim HJ, Douek M, Gutierrez A, Patel M, et al. RECIST 1.1 Target Lesion Categorical Response in Metastatic Renal Cell Carcinoma: A Comparison of Conventional versus Volumetric Assessment. Radiol Imaging Cancer. 2023;5(5):e220166. https://doi.org/10.1148/rycan.220166.

24. Dahm IC, Kolb M, Altmann S, Nikolaou K, Gatidis S, Othman AE, et al. Reliability of Automated RECIST 1.1 and Volumetric RECIST Target Lesion Response Evaluation in Follow-Up CT-A Multi-Center, Multi-Observer Reading Study. Cancers (Basel). 2024;16(23). https://doi.org/10.3390/cancers16234009.

25. Funingana IG, Piyatissa P, Reinius M, McCague C, Basu B, Sala E. Radiomic and Volumetric Measurements as Clinical Trial Endpoints-A Comprehensive Review. Cancers (Basel). 2022;14(20). https://doi.org/10.3390/cancers14205076.

26. Beaumont H, Cantini L, Saini KS, Faye N, Gill R, Iannessi A. What are RECIST 1.1 progressions made of? Variability in double-read oncology trials. Eur Radiol. 2026;36(7):6159–6169. https://doi.org/10.1007/s00330-025-12234-4.

27. Tareco Bucho TM, Pascucci L, Mancò G, Taraschi A, Macchia MC, Tinari N, et al. RECIST progression: Patterns among target, non-target, non-measurable and new lesions progression. Eur J Radiol. 2025;186:112038. https://doi.org/10.1016/j.ejrad.2025.112038.

28. Rauwerdink DJW, Molina G, Frederick DT, Sharova T, van der Hage J, Cohen S, et al. Mixed Response to Immunotherapy in Patients with Metastatic Melanoma. Ann Surg Oncol. 2020;27(9):3488–3497. https://doi.org/10.1245/s10434-020-08657-6.

29. Jan Willem Rauwerdink D, van Persijn van Meerten E, van der Hage J, Kapiteijn E. Management of heterogeneous tumor response patterns to immunotherapy in patients with metastatic melanoma. Melanoma Res. 2022;32(1):45–54. https://doi.org/10.1097/CMR.0000000000000794.

30. Versluis JM, Hoefsmit EP, Shehwana H, Dimitriadis P, Sanders J, Broeks A, et al. Tumor characteristics of dissociated response to immune checkpoint inhibition in advanced melanoma. Cancer Immunol Immunother. 2024;73(2):28. https://doi.org/10.1007/s00262-023-03581-6.

31. Guan Y, Cui Y, Gong Y, Liang X, Han X, Chen Y, et al. Dissociated response and treatment outcome with immune checkpoint blockade in advanced cancer. Sci Rep. 2024;14(1):32147. https://doi.org/10.1038/s41598-024-84009-8.

32. Morinaga T, Inozume T, Kawazu M, Ueda Y, Sax N, Yamashita K, et al. Mixed Response to Cancer Immunotherapy is Driven by Intratumor Heterogeneity and Differential Interlesion Immune Infiltration. Cancer Res Commun. 2022;2(7):739–753. https://doi.org/10.1158/2767-9764.CRC-22-0050.

33. Beck L, Harel M, Yu S, Markovits E, Boursi B, Markel G, et al. Clinical Proteomics of Metastatic Melanoma Reveals Profiles of Organ Specificity and Treatment Resistance. Clin Cancer Res. 2021;27(7):2074–2086. https://doi.org/10.1158/1078-0432.CCR-20-3752.

34. Conway JW, Rawson RV, Lo S, Ahmed T, Vergara IA, Gide TN, et al. Unveiling the tumor immune microenvironment of organ-specific melanoma metastatic sites. J Immunother Cancer. 2022;10(9). https://doi.org/10.1136/jitc-2022-004884.

35. Szeitz B, Hagemeijer YP, Pahi ZG, Ujfaludi Z, Kuras M, Rodriguez J, et al. Distant metastases of melanoma exhibit varying extent of intrapatient proteogenomic heterogeneity. Clin Transl Med. 2025;15(10):e70477. https://doi.org/10.1002/ctm2.70477.

36. Chen DT, Chan W, Thompson ZJ, Thapa R, Beg AA, Saltos AN, et al. Utilization of target lesion heterogeneity for treatment efficacy assessment in late stage lung cancer. PLoS One. 2021;16(7):e0252041. https://doi.org/10.1371/journal.pone.0252041.

37. Topp BG, Thiagarajan K, De Alwis DP, Snyder A, Hellmann MD. Lesion-level heterogeneity of radiologic progression in patients treated with pembrolizumab. Ann Oncol. 2021;32(12):1618–1625. https://doi.org/10.1016/j.annonc.2021.09.006.

38. Pan J, Wu J, Wang B, Zhu B, Liu X, Gan H, et al. Interlesional response heterogeneity is associated with the prognosis of abiraterone treatment in metastatic castration-resistant prostate cancer. Med. 2024;5(12):1475–1484.e3. https://doi.org/10.1016/j.medj.2024.07.020.

39. Ramakrishnan D, Brüningk SC, von Reppert M, Memon F, Maleki N, Aneja S, et al. Comparison of Volumetric and 2D Measurements and Longitudinal Trajectories in the Response Assessment of BRAF V600E-Mutant Pediatric Gliomas in the Pacific Pediatric Neuro-Oncology Consortium Clinical Trial. AJNR Am J Neuroradiol. 2024;45(4):475–482. https://doi.org/10.3174/ajnr.A8189.

40. Henry T, Sun R, Lerousseau M, Estienne T, Robert C, Besse B, et al. Investigation of radiomics based intra-patient inter-tumor heterogeneity and the impact of tumor subsampling strategies. Sci Rep. 2022;12(1):17244. https://doi.org/10.1038/s41598-022-20931-z.

41. Zimmermann M, Kuhl CK, Engelke H, Bettermann G, Keil S. CT-based whole-body tumor volumetry versus RECIST 1.1: Feasibility and implications for inter-reader variability. Eur J Radiol. 2021;135:109514. https://doi.org/10.1016/j.ejrad.2020.109514.

42. Yan K, Wang X, Lu L, Summers RM. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imaging (Bellingham). 2018;5(3):036501. https://doi.org/10.1117/1.JMI.5.3.036501.

43. Shiyam Sundar LK, Yu J, Muzik O, Kulterer OC, Fueger B, Kifjak D, et al. Fully Automated, Semantic Segmentation of Whole-Body 18F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence. J Nucl Med. 2022;63(12):1941–1948. https://doi.org/10.2967/jnumed.122.264063.

44. Leung KH, Rowe SP, Sadaghiani MS, Leal JP, Mena E, Choyke PL, et al. Deep Semisupervised Transfer Learning for Fully Automated Whole-Body Tumor Quantification and Prognosis of Cancer on PET/CT. J Nucl Med. 2024;65(4):643–650. https://doi.org/10.2967/jnumed.123.267048.

45. Joskowicz L, Szeskin A, Rochman S, Dodi A, Lederman R, Fruchtman-Brot H, et al. Follow-up of liver metastases: a comparison of deep learning and RECIST 1.1. Eur Radiol. 2023;33(12):9320–9327. https://doi.org/10.1007/s00330-023-09926-0.

46. Küstner T, Peisen F, Gatidis S, Wagner A, Megne O, Othman A, et al. Longitudinal-CT. University of Tübingen; 2025. https://doi.org/10.57754/FDAT.75kj1-64747.

Tables and Table Captions

Table 1. Target–all and direct target–non-target comparisons

Comparison	n	ρ	MAE	20%	30%	50%
TD vs AD, full cohort	264	0.873	0.171	16.29%	18.94%	17.80%
TD vs AD, non-target eligible	228	0.845	0.198	18.86%	21.93%	20.61%
TD vs NTD	228	0.659	0.321	29.39%	34.65%	33.77%
TV vs AV, full cohort	264	0.894	0.548	8.33%	9.09%	11.36%
TV vs AV, non-target eligible	229	0.875	0.632	9.61%	10.48%	13.10%
TV vs NTV	229	0.678	1.357	21.40%	23.58%	27.95%

Abbreviations: TD, target-lesion diameter change; AD, all-lesion diameter change; NTD, non-target tracked-lesion diameter change; TV, target-lesion volume change; AV, all-lesion volume change; NTV, non-target tracked-lesion volume change; ρ, Spearman rank correlation coefficient; MAE, mean absolute error. Threshold columns report response-direction discordance.

Table 2. All-lesion underrepresentation and direct non-target mixed-response capture

Threshold	All-lesion mixed	Target also mixed	Underrepresented, n (%; 95% CI)	NTD mixed	Target also mixed, n (%; 95% CI)	NTV mixed / target also mixed, n (%; 95% CI)
20%	73	17	56 (76.7; 65.8–84.9)	55	7 (12.7; 6.3–24.0)	84 / 17 (20.2; 13.0–30.0)
30%	58	11	47 (81.0; 69.1–89.1)	45	4 (8.9; 3.5–20.7)	78 / 16 (20.5; 13.0–30.8)
50%	40	6	34 (85.0; 70.9–92.9)	34	2 (5.9; 1.6–19.1)	61 / 8 (13.1; 6.8–23.8)

Abbreviations: NTD, non-target tracked-lesion diameter mixed response; NTV, non-target tracked-lesion volume mixed response; CI, confidence interval. Diameter analyses included the 2-mm absolute-change safeguard.

Table 3. Deterministic selection strategies in the common cohort

Strategy	Diameter ρ	Diameter MAE	Volume ρ	Volume MAE	Diameter discordance (20/30/50%)
Radiologist designated	0.871	0.161	0.892	0.413	15.6/18.8/17.6%
Largest diameter	0.915	0.146	0.953	0.292	16.8/17.2/16.0%
Organ-balanced diameter	0.908	0.141	0.948	0.282	16.8/15.2/19.2%
Largest volume	0.914	0.148	0.953	0.296	17.6/18.8/16.0%

Abbreviations: ρ, Spearman rank correlation coefficient; MAE, mean absolute error.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.zip