

How errors from prompt-assisted longitudinal tumour segmentation change RECIST 1.1 target-lesion response categories: a retrospective out-of-fold analysis

Souraj Adhikary

souraj.adhikary@student.jade-hs.de

Jade Hochschule Wilhelmshaven

Negar Chabi

Jade Hochschule Wilhelmshaven

Andre Mastmeyer

Jade Hochschule Wilhelmshaven

Research Article

Keywords: Response Evaluation Criteria in Solid Tumors, Tomography (X-ray computed), Deep learning, Image segmentation, Reproducibility of results

Posted Date: August 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10551828/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 How errors from prompt-assisted longitudinal
2 tumour segmentation change RECIST 1.1
3 target-lesion response categories: a retrospective
4 out-of-fold analysis

5 Souraj Adhikary^{1*}, Negar Chabi¹ and André Mastmeyer¹
6 ^{1*}Department of Engineering Sciences, Jade Hochschule Wilhelmshaven,
7 Friedrich-Paffrath-Straße 101, 26389, Wilhelmshaven, Germany.

8 *Corresponding author(s). E-mail(s):
9 souraj.adhikary@student.jade-hs.de;
10 Contributing authors: negar.chabi@jade-hs.de;
11 andre.mastmeyer@jade-hs.de;

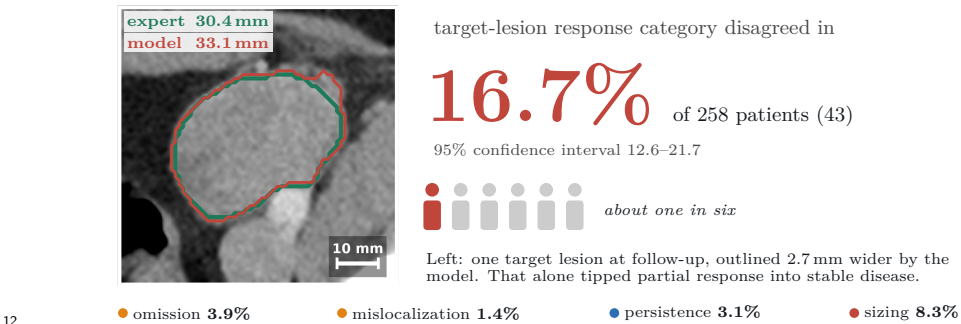


Fig. 1 Study at a glance. Each patient was segmented by the cross-validation fold that excluded that patient, and the expert follow-up mask was withheld. Percentages are Shapley contributions to the 16.7% rate; sizing was largest but not statistically separated from omission and mislocalization combined (5.2%), so no channel clearly dominated. Scope: the target-lesion component of RECIST 1.1 only, measured from masks, in an internal out-of-fold evaluation of one model (LongiSeg) with a correct expert baseline; overall RECIST 1.1 response was not assessed. Computed tomography excerpt (expert outline green, model red) from the autoPET/CT longitudinal dataset, version 2 (CC BY 4.0).

Abstract

Background: RECIST 1.1 requires clinicians to compare target-lesion measurements across examinations. Automated segmentation may support this task, but omission, tracking, persistence, and measurement errors need not affect response equally. We asked how often these errors changed target-lesion response and whether changes favoured apparent benefit.

Methods: We retrospectively analysed 258 response-evaluable patients from a public longitudinal whole-body CT cohort with expert per-lesion masks. LongiSeg received the baseline CT, expert baseline mask and an interior location click, plus the follow-up CT and a location click; the expert follow-up mask was withheld. We compared a RECIST 1.1 target-lesion response based on expert masks with one based on expert baseline and model follow-up masks. Each patient was processed by a held-out-fold checkpoint not fitted to that patient. Shapley attribution divided disagreements among prompted omission, tracking mislocalization, false persistence, and measurement (sizing) error. Omission and mislocalization were combined as detection/localization. We estimated disagreement with a Wilson CI, channel contributions with patient-bootstrap CIs, and tested the plausibility of sizing effects in a follow-up-only simulation.

Results: Categories disagreed in 43/258 patients (16.7%; Wilson 95% CI 12.6–21.7; Cohen $\kappa = 0.76$). Sizing contributed 8.3% (95% CI 5.0–11.8), detection/localization 5.2% (2.7–7.9), and false persistence 3.1% (1.2–5.4). Sizing exceeded persistence and omission separately, but not detection/localization combined (+3.1 percentage points, 95% CI -1.2 to 7.4); no channel clearly dominated. Multiple errors occurred in 13/43 disagreements (30%). Post hoc, 28 disagreements moved toward apparent benefit and 15 toward harm, but this pattern changed with target selection and did not test objective-response rate. Resampled sizing errors gave a 9.3% mean flip rate (central 95% simulation interval 6.2–12.4), which encompassed the 8.3% sizing contribution. Flip rates ranged from 15.5% to 18.6% across sensitivity analyses and from 15.0% to 18.4% across leave-one-fold-out subsets.

Conclusions: Segmentation errors changed target-lesion response in *about one in six* patients. Measurement error was important but not clearly dominant. These findings concern follow-up errors from one prompt-assisted model, conditional on a correct expert baseline, in an internal out-of-fold evaluation. Independent confirmation is needed, especially for disagreement direction.

Keywords: Response Evaluation Criteria in Solid Tumors, Tomography (X-ray computed), Deep learning, Image segmentation, Reproducibility of results

1 Background

RECIST 1.1 is the reference standard for assessing solid-tumour response in oncology trials and is increasingly used in routine reporting [1, 2]. Clinicians measure selected target lesions at baseline and follow-up, compare their summed diameters, and assign

complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD). These categories define trial endpoints and can influence management. Measurement matters because even small differences between readers can change the category [3, 4].

Automated segmentation could make this process more reproducible and scalable [5–7]. Yet a small boundary error can cross a RECIST threshold. Most safety discussions focus on missed lesions, which can understate tumour burden. Detection is therefore the error most often audited. Semi-automatic tools are commonly viewed as safer because the user identifies the lesions to measure [8], but these tools can still track the wrong structure or measure a known lesion inaccurately.

Follow-up segmentation can fail in four ways: omit a lesion, track it to the wrong location, retain it after it has resolved, or measure it inaccurately. Prior studies have measured reader variability in manual and semi-automatic RECIST [8–10] and the overall reliability of automated response pipelines [6, 11]. To our knowledge, however, no study has used patient-level held-out-fold predictions on real follow-up CT to determine how much each error type changes the response category. We therefore asked which errors cause category disagreement and whether disagreements tend to make response look better [12].

We studied LongiSeg, a prompt-assisted longitudinal segmentation model, on a public whole-body CT cohort with expert outlines for each lesion. The model received the expert baseline lesion and a challenge-supplied location at follow-up, but not the expert follow-up outline. This is a favourable assisted-tracking task, not autonomous whole-body detection. Each patient was evaluated with a checkpoint that had not been fitted to that patient’s scans. We measured how often model errors changed the target-lesion response category. We then divided each disagreement among the four error types and tested whether sizing errors of the observed magnitude could plausibly explain the sizing contribution. Fig. 1 summarizes the design and the main result.

81 2 Methods

82 2.1 Study design and data

83 We retrospectively analysed the public autoPET/CT longitudinal whole-body CT
84 dataset [13, 14]. It contains baseline and follow-up examinations acquired during
85 systemic therapy from patients with metastatic melanoma at one centre (Univer-
86 sity Hospital Tübingen). Expert readers outlined each lesion separately, assigned an
87 anatomical type, and linked lesions across timepoints; several baseline labels could
88 map to one follow-up label when lesions coalesced.¹

89 The LongiSeg development cohort contained 300 patients: 181 male and 119 female
90 according to the dataset descriptor [14]. We analysed the version 1 images and masks;
91 version 2 supplied only explicit topology fields used to validate coalescence [15]. The
92 processed files did not contain individual age or sex, so these variables could not be
93 reported for the 258 analysed patients and age- or sex-stratified subgroup analyses
94 were not possible. The 300-patient sex distribution is provided only as source-cohort
95 context. A recovery audit of seven initially missing downloads is described below.

96 We followed CLAIM 2024 for artificial-intelligence imaging studies and STROBE
97 for observational studies. Every algorithm behind the reported numbers is specified in
98 pseudocode in [Appendix A](#).

99 2.2 RECIST 1.1 target-lesion calculation

100 At baseline, RECIST 1.1 selects at most five measurable target lesions in total and no
101 more than two per organ [1]. Non-nodal CT targets are measured by longest diameter
102 and must be at least 10 mm; pathological lymph-node targets are measured by short
103 axis² and must be at least 15 mm. The selected measurements are summed as the sum

¹Coalesced: two or more separate lesions have grown together by follow-up and can no longer be measured apart. The single resulting mass is called a conglomerate.

²Short axis: the shorter of the two perpendicular diameters of a lymph node in the imaging plane. Nodes are enlarged mainly across their short axis, so it separates disease from normal nodes better than the longest diameter.

of diameters (SLD). At follow-up, complete response (CR) requires disappearance of non-nodal targets and reduction of target nodes below 10 mm. Partial response (PR) requires an SLD decrease of at least 30% from baseline. Progressive disease (PD) requires an SLD increase of at least 20% from the on-study nadir³ and an absolute increase of at least 5 mm. Stable disease (SD) is neither PR nor PD.

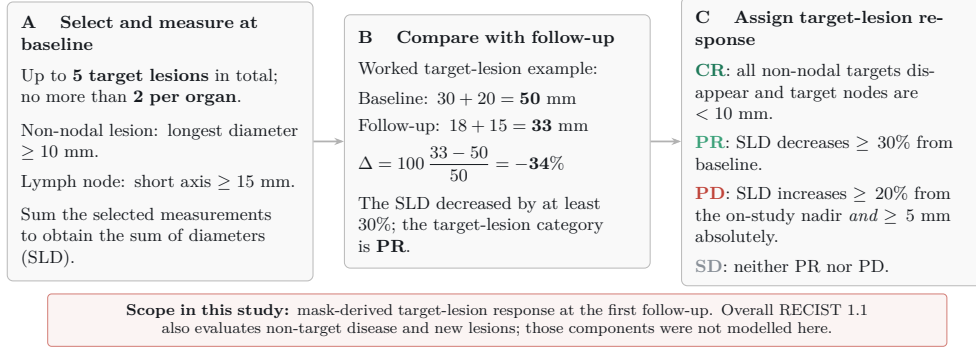


Fig. 2 RECIST 1.1 target-lesion calculation. The schematic summarizes baseline selection and measurement, a worked SLD example, and target-lesion response thresholds. Overall RECIST 1.1 also evaluates non-target disease and new lesions; this study models only the target-lesion component. CR denotes complete response; PR, partial response; SD, stable disease; PD, progressive disease; and SLD, sum of diameters.

2.3 Study-specific target selection and endpoint

Our endpoint was a mask-derived, deliberately *stripped-down* version of RECIST 1.1: the target-lesion component only, not overall RECIST response. We selected targets from expert baseline masks using a mask-based approximation of RECIST 1.1 [1]. Measurable lesions were ranked from largest to smallest, with at most two per dataset lesion type and five in total.

The dataset’s lesion types include specific sites, such as liver, lung, adrenal and spleen, and broad groups, such as lymph node, soft tissue/skin and other. The two-per-type rule matches the RECIST two-per-organ rule for organ-specific types but

³Nadir: the smallest SLD recorded so far on study. Progression is judged against this smallest value, not against baseline, so growth after an initial shrinkage is still detected.

118 groups the broader types more coarsely. Seventeen of the 683 selected targets (2.5%)
119 carried the dataset’s *unclear* type; these shared one type bucket for the two-per-type
120 cap. RECIST permits lytic⁴ or mixed lytic–blastic bone lesions with a measurable soft-
121 tissue component; because the dataset does not identify that component or osseous
122 phenotype,⁵ we excluded all bone-typed lesions from the primary analysis.

123 The SLD summed the longest diameter of each non-nodal target and the short axis
124 of each nodal target. For nodes, we used the largest-area axial slice and measured the
125 greatest width perpendicular to its longest in-slice caliper. A node always contributed
126 its short axis to the SLD; the < 10 mm rule was used only when testing for CR.

127 We compared baseline with the first follow-up and applied the target-lesion rules
128 in Fig. 2. Some examinations contained overlapping body-region CT volumes; for both
129 reference and model, we measured each target across all volumes and retained its
130 largest measurement so that a lesion in a later volume was not treated as absent.

131 For true coalescence, we applied the RECIST conglomerate rule [1]. The baseline
132 SLD retained the selected source-lesion diameters; the surviving follow-up conglomerate
133 was measured once. The reference used the surviving label, whereas the model
134 used the union of predicted labels in the full merge group so that a label change was
135 not mistaken for disappearance. Nineteen merge groups involved 28 selected sources
136 in 17 patients; all matched the version 2 topology audit. The 683 selected baseline
137 targets therefore formed 674 follow-up assessment units.

138 New lesions were excluded because LongiSeg tracks prompted, known lesions rather
139 than searching the whole body; expert new-lesion status was retained only as a descrip-
140 tive stratum. We also did not assess non-target disease. The comparator was the expert
141 mask, not an independent radiologist read. The endpoint is therefore limited to the
142 target-lesion component of RECIST, measured identically for reference and model.

⁴Lytic: a bone metastasis that destroys bone, appearing on CT as a dark, low-density hole. A blastic lesion does the opposite and lays down dense new bone; a mixed lytic–blastic lesion does both.

⁵Osseous phenotype: which of these forms a given bone lesion takes. RECIST measures only the soft-tissue mass some bone lesions carry, never the bone change itself, so the phenotype decides whether a bone lesion is measurable at all.

2.4 Segmentation model and patient-level out-of-fold

evaluation

LongiSeg [16] is a prompt-assisted longitudinal tumour-segmentation model with an nnU-Net residual-encoder backbone [17]. We used the public `autoPET` branch at commit `ee5296ee4b0ebc4f56aa916c34215edb9629dace` [16]. For each lesion, the model’s inputs and output were:

Inputs: baseline CT scan + expert baseline lesion outline (mask) + baseline location click
+ follow-up CT scan + follow-up location click
→ **LongiSeg model trained without that patient**
→ **Output: predicted follow-up lesion outline (mask)**

The expert outline of the lesion at follow-up was not an input. It was withheld and used only as the reference for evaluating the predicted outline. In other words, LongiSeg was shown exactly how the lesion was outlined at baseline and approximately where to look at follow-up, then had to outline it at follow-up. Fig. 3 summarizes this workflow and the error taxonomy used in this study.

For each lesion the released pipeline worked on a fixed crop of $112 \times 224 \times 224$ voxels. The crop held five channels: the follow-up CT, the baseline CT, the expert baseline outline, and a Gaussian blob marking each of the two clicks. The follow-up click positioned the crop in the new scan, and the baseline click aligned the matching baseline crop. Because the expert baseline outline already gave the lesion boundary, the baseline click was an architectural input rather than a second clinical delineation. The network produced the follow-up outline only.

The wrapper processed all linked challenge lesions across matching baseline–follow-up volume pairs and merged labelled outputs before measurement.

Each lesion came with two candidate follow-up locations: one propagated from the baseline lesion by image registration (`cog_propagated`), and one taken from the

169 follow-up annotation itself (`cog_fu`). The challenge uses the propagated location where
170 it exists and the annotated location otherwise; of the 683 selected baseline sources,
171 678 prompts (99.3%) were propagated and five (0.7%) were fallbacks. The annotated
172 location is the more informative of the two, because it points at where the lesion
173 actually is rather than where registration predicts it, so we also report the disagreement
174 rate with the fallback-affected patients excluded. Either way the model is told roughly
175 where to look, which is more favourable support than autonomous detection: an empty
176 or displaced output is a prompted tracking or segmentation failure, not a detection
177 miss [18, 19]. A propagated location may still fall outside a lesion that has moved,
178 changed size, or resolved.

179 LongiSeg was developed with five-fold cross-validation on these same 300 patients,
180 and its published configuration was selected on this cohort [16]. We reconstructed the
181 group-preserving split and used only each patient’s held-out-fold checkpoint,⁶ without
182 ensembling. Thus, the evaluated patient’s scans were excluded from weight fitting, but
183 configuration-selection optimism can remain. Median held-out-fold per-case Dice⁷ was
184 0.77 (interquartile range 0.46–0.85).

185 Two inference-only corrections preceded the final analysis: the wrapper recovered
186 predictions for 133 lesions in 29 patients whose baseline masks were outside volume 00,
187 and a one-voxel upper-bound guard corrected one valid edge click. Neither changed
188 model weights. Every prompted target in the analysed cohort consequently had a held-
189 out-fold prediction. Baseline measurements came from expert masks; follow-up model
190 measurements used predicted masks and the same caliper rules.

⁶Held-out-fold (out-of-fold) checkpoint: cross-validation splits the cohort into five folds and trains five models, each on four folds. A patient is always segmented by the one model whose training excluded that patient.

⁷Dice: the standard overlap score between two masks, from 0 for no shared voxels to 1 for identical masks.

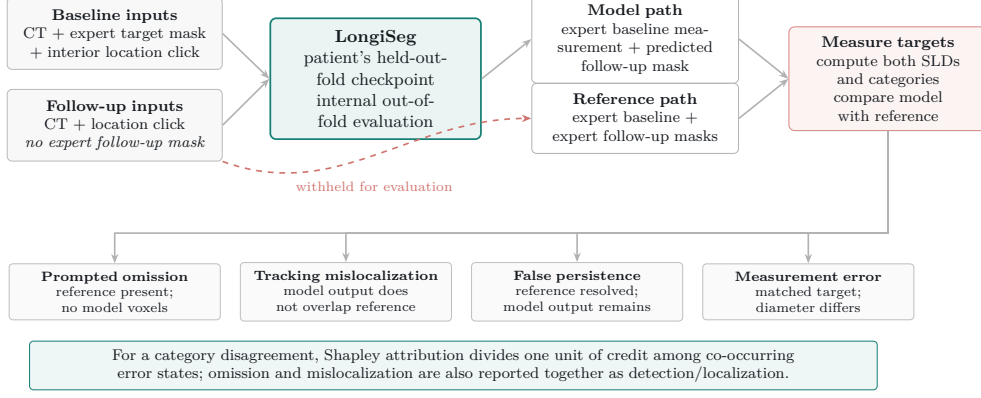


Fig. 3 Study workflow and error taxonomy. LongiSeg received baseline and follow-up computed tomography (CT), the expert baseline mask, and location clicks; the expert follow-up mask was withheld for evaluation. Expert and model follow-up masks produced parallel target-lesion SLDs and response categories. SLD denotes sum of diameters.

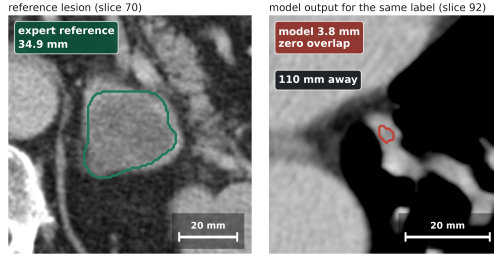
2.5 Mechanism decomposition

We compared the model with the expert reference for each follow-up assessment unit. A single target formed one unit; targets in the same true coalescence shared one unit and one error state:

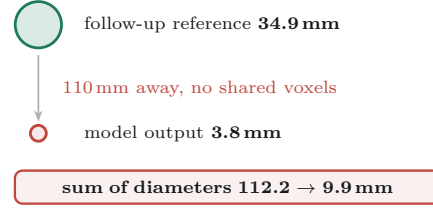
1. **Prompted omission:** the reference unit was present, but the model produced no voxels for any label in its full merge group.
2. **Tracking mislocalization:** the model produced the expected label or a merge-group label, but its voxels did not overlap the reference unit.
3. **False persistence:** the reference unit had resolved, but model output remained—a model-side false positive for a previously known target.
4. **Measurement (sizing) error:** the model and reference overlapped, but their measured diameters differed.

Sizing was therefore limited to lesions the model located. A non-overlapping prediction is a localization error, not a measurement error. We combined omission and mislocalization as the *detection/localization* channel. Fig. 4 shows a real example of each of the last three states; prompted omission is shown in Fig. 9.

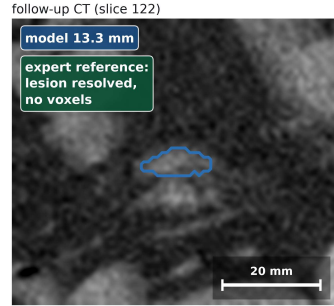
(A) Tracking mislocalization



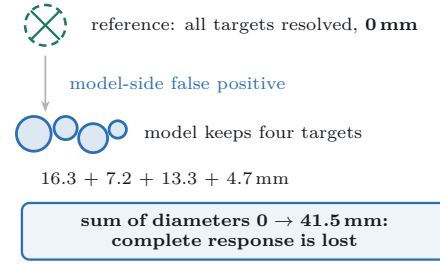
reference SD $\rightarrow$ model PR (apparent benefit)



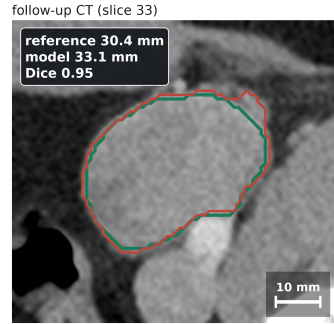
(B) False persistence



reference CR $\rightarrow$ model PR (apparent harm)



(C) Measurement (sizing) error



reference PR $\rightarrow$ model SD (apparent harm)

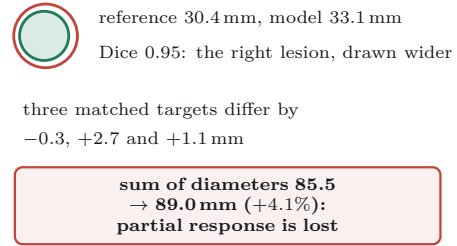


Fig. 4 Three of the four error states in real data. The fourth state, prompted omission, is shown in Fig. 9. Each row pairs a follow-up computed tomography (CT) slice with a schematic of the same case. Green marks the expert reference outline and red or blue the model outline. **(A)** Tracking mislocalization: the model produced output for the prompted label 110 mm away from the reference lesion, with no shared voxels. **(B)** False persistence: every reference target had resolved, yet the model still outlined four of them. **(C)** Measurement (sizing) error: the model found the right lesion (Dice 0.95) but drew it wider, and the summed diameters crossed a response threshold. Millimetre values are the analysed diameters for these patients. No patient's category changed through mislocalization alone. Images are from autoPET/CT longitudinal dataset version 2 (CC BY 4.0; DOI: 10.57754/FDAT.75kj1-64747) [15], cropped and annotated by the authors. CR denotes complete response; PR, partial response; and SD, stable disease.

207 The primary definition of a spatial match was any voxel overlap ($\text{Dice} > 0$); sen-
 208 sitivity analyses used $\text{Dice} > 0.05$ and > 0.10 . This threshold affects only the error
 209 label. The response category always uses the measurements from the model’s actual
 210 mask, so the disagreement rate does not change with the matching threshold.

211 When a patient had more than one error type, the assigned contribution could
 212 depend on which error was corrected first. We avoided this ordering problem by apply-
 213 ing corrections in every possible order and using Shapley attribution⁸ to average how
 214 much correcting each error type reduced category disagreement across those orders.
 215 For each correction order, we scored the patient as 1 while the model and reference
 216 categories still differed and 0 once they agreed. Correcting omission, mislocalization or
 217 sizing replaced the affected follow-up measurement with the reference value; correct-
 218 ing false persistence removed model output for a resolved lesion. Error types absent
 219 from a patient received zero contribution.

220 In one disagreement, the model category was partial response (PR) and the ref-
 221 erence category was stable disease (SD). Neither correcting omission nor correcting
 222 mislocalization alone resolved it; correcting both did, so Shapley attribution assigned
 223 0.5 to each, and the other two error types were null players with zero contribution
 224 (Table 1).

Table 1 Worked example of Shapley attribution in one disagreement (reference stable disease, model partial response). Correcting omission alone or mislocalization alone left the categories different; correcting both restored stable disease, so each received a contribution of 0.5.

Target	Reference (mm)	Model (mm)	Dice	Error state
Target 1	34.9	3.8	0.00	Tracking mislocalization
Target 2	26.2	0.0	—	Prompted omission
Target 3	25.4	0.0	—	Prompted omission
Target 4	25.7	6.1	0.00	Tracking mislocalization

⁸Shapley attribution: a credit-sharing rule from cooperative game theory. Each error type is credited with its average contribution over every possible correction order, so no error is favoured by being corrected first.

225 With four possible error types, we evaluated all 16 correction subsets and all 24
226 correction orders. No correction increased disagreement in any of the 43 patients whose
227 model and reference categories differed, and correcting all errors reproduced the ref-
228 erence category. Contributions were therefore non-negative and summed to one per
229 disagreement. We summed these patient-level contributions and divided by 258 to
230 obtain cohort percentages. The attribution is the standard Shapley value of this cor-
231 rection game [20]. This design follows recommendations for avoiding metric-related
232 pitfalls in image-analysis validation [21, 22].

233 2.6 Synthetic sizing control and statistics

234 We used three follow-up-only simulations to ask whether measurement error alone
235 could cause a realistic number of category flips. Baseline measurements remained fixed
236 at the expert reference, as in the main analysis, and resolved targets remained absent.
237 First, we multiplied each present follow-up diameter by factors from 0.60 to 1.40 in
238 steps of 0.05. Second, we moved its boundary uniformly by 0, ± 0.15 , ± 0.30 , ± 0.60 ,
239 ± 1.0 , or ± 1.5 mm; the diameter changes by twice that distance. Third, we sampled
240 with replacement from the relative diameter errors in spatially matched held-out-fold
241 predictions and applied the sampled error to the reference diameter. As a secondary
242 demonstration, we applied the same multiplicative factor at baseline and follow-up to
243 show how common error at both timepoints cancels in the ratio.

244 We report the overall disagreement proportion with a Wilson score 95% CI.⁹
245 Channel percentages and their differences were computed from unrounded patient-
246 level contributions and rounded only for display, so displayed components need not
247 sum exactly. Channel proportions use patient-level nonparametric bootstrap 95%

⁹Wilson score interval: a confidence interval for a proportion that stays inside 0–100% and remains reliable for small counts, unlike the textbook normal-approximation interval.

248 CIs¹⁰ (10 000 resamples, fixed seed). Channel differences use a paired bootstrap. The
249 direction analysis uses an exact binomial test.

250 The empirical sizing simulation used 2 000 Monte Carlo replicates with random
251 seed 20260711. We report the mean flip rate and the 2.5th and 97.5th percentiles as a
252 central 95% simulation interval. This is not a confidence interval from an independent
253 sample; it describes variation conditional on this cohort and its observed error pool.
254 Errors were sampled independently for each follow-up unit, so the simulation does not
255 model within-patient correlation and remains model-specific. The multiplicative and
256 boundary simulations are model-agnostic.

257 Sensitivity analyses varied the RECIST thresholds, the nodal measurement (caliper
258 versus a principal component analysis [PCA] minor-axis estimate), the spatial-match
259 threshold (Dice > 0 , > 0.05 , or > 0.10), and the attribution order. On the same
260 258 patients, we also allowed one or five targets per lesion type while retaining the
261 five-target total cap. A deliberately inclusive skeletal upper bound treated every size-
262 eligible bone mask as measurable. Finally, we audited the prompt source for all 683
263 selected sources and recalculated the main rate after excluding patients with any
264 follow-up-centroid (`cog_fu`) fallback prompt. All analyses were post hoc; none was
265 pre-registered.

266 Generative AI was used only for light language editing and grammar correction.

267 The authors reviewed the final text and take full responsibility for the manuscript.

268 3 Results

269 3.1 Cohort and headline rate

270 The recovery audit restored all seven initially missing downloads, so all 300 devel-
271 opment patients were present and had an assigned held-out-fold checkpoint. One

¹⁰Patient-level nonparametric bootstrap: the cohort is resampled with replacement, whole patients at a time, and the analysis is repeated on each resample. The spread of the results gives the interval, with no distributional assumption.

272 patient remained technically non-evaluable because six registration-propagated follow-
 273 up clicks fell outside a resampled image volume; 41 others had no measurable baseline
 274 target. The final analysis therefore included 258 patients (Fig. 5; Table 2).

275 Twenty-eight patients had an expert-defined new lesion at follow-up; this was
 276 reported only as a descriptive stratum. The model and expert-mask reference agreed
 277 on the target-lesion category in 215/258 patients (83.3%; Cohen $\kappa = 0.757$). They dis-
 278 agreed in 43/258 (16.7%; Wilson 95% CI 12.6–21.7). Almost all disagreements moved
 279 by one adjacent category (Fig. 6); none moved between CR and PD.

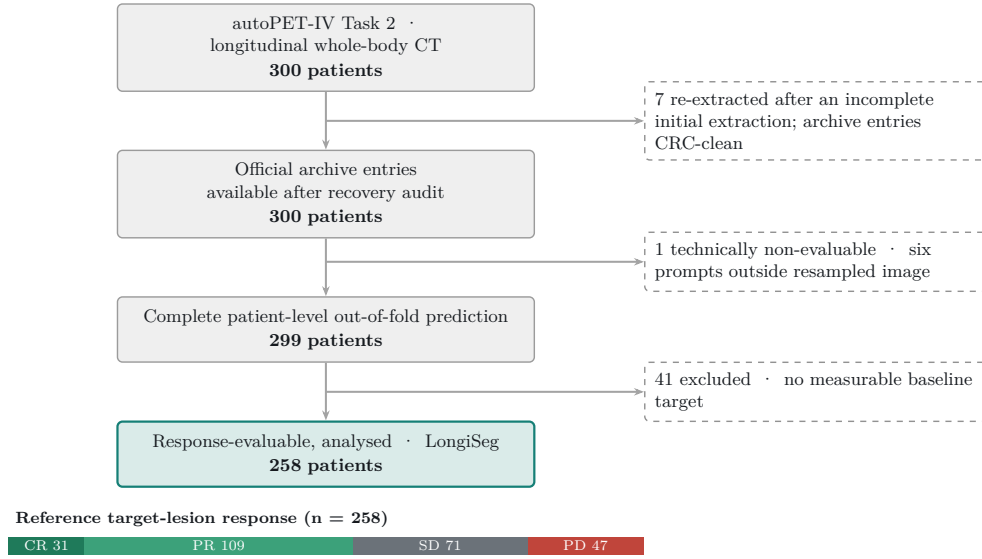


Fig. 5 Study cohort. Flow from the 300-patient autoPET-IV Task 2 cohort to 258 response-evaluable patients with held-out-fold predictions and their expert-mask target-lesion response distribution. CR denotes complete response; PR, partial response; SD, stable disease; and PD, progressive disease.

280 3.2 No error channel clearly dominated

281 Sizing error among spatially matched units accounted for 8.3% of the cohort (95% CI
 282 5.0–11.8). Detection/localization accounted for 5.2% (2.7–7.9): 3.9% from prompted

		Model (out-of-fold)			
		CR	PR	SD	PD
Reference (expert mask)	CR	23	8		
	PR	4	102	3	
	SD	1	12	54	4
	PD		3	8	36

43/258 disagree (16.7%); off-diagonal (red) are the flips

Fig. 6 Expert reference versus model target-lesion response. The 43 off-diagonal cases (red) are category disagreements; almost all moved by one category. CR denotes complete response; PR, partial response; SD, stable disease; and PD, progressive disease.

Table 2 Study cohort characteristics.

Characteristic	Value
Patients analysed (of 300 in development cohort)	258
Disease; centre	Metastatic melanoma; single centre (Univ. Tübingen)
Age and sex (258 analysed patients)	Not available in the processed analysis files
Sex (300-patient development cohort only, per descriptor)	181 male, 119 female
Timepoints compared	Baseline → first follow-up
Baseline target lesions (total; per patient)	683; median 2 (range 1–5)
Target lesion type (n)	lymph node 155, soft tissue/skin 146, liver 118, lung 116, others 81, adrenals 33, unclear 17, spleen 10, CNS 3, kidney 2, heart 2
Expert-mask target-lesion category (CR / PR / SD / PD)	31 / 109 / 71 / 47
Model target-lesion category (CR / PR / SD / PD)	28 / 125 / 65 / 40
Patients with a ground-truth new lesion (stratum)	28 (10.9%)
Follow-up assessment units (model vs. reference)	674
Both present	502 (483 spatially matched, 19 mislocalized)
Of the 483 matched units	470 differed in diameter, 13 identical
Omitted by the model	28
Model-only persistence	39
Absent in both	105

Note: The expert category uses expert masks at both timepoints. The model category uses the same expert baseline mask and the predicted follow-up mask; the expert follow-up mask was not a model input. The 683 selected baseline sources formed 674 follow-up assessment units. CR denotes complete response; PR, partial response; SD, stable disease; and PD, progressive disease.

omission (1.7–6.4) and 1.4% from tracking mislocalization (0.2–2.9). False persistence accounted for 3.1% (1.2–5.4) (Fig. 7).

Sizing had the largest estimate. It exceeded persistence by +5.2 percentage points (95% CI 1.4–9.3) and omission by +4.5 points (0.4–8.5). However, it was not separated from detection/localization combined: +3.1 points (−1.2 to 7.4). Persistence was also not separated from detection/localization: −2.1 points (−5.6 to 1.2). These results identify sizing as an important contributor, but not as an unequivocally dominant one.

Thirty disagreements involved one error state, 12 involved two, and one involved three, so 13/43 (30%) contained more than one error. Each disagreement contributed one unit of attribution, split among the error states it contained, and the channel

293 totals are expressed over all 258 patients. Each percentage is therefore that channel's
 294 share of the cohort's disagreement burden, not the chance that one such error will
 295 change an individual patient's category.

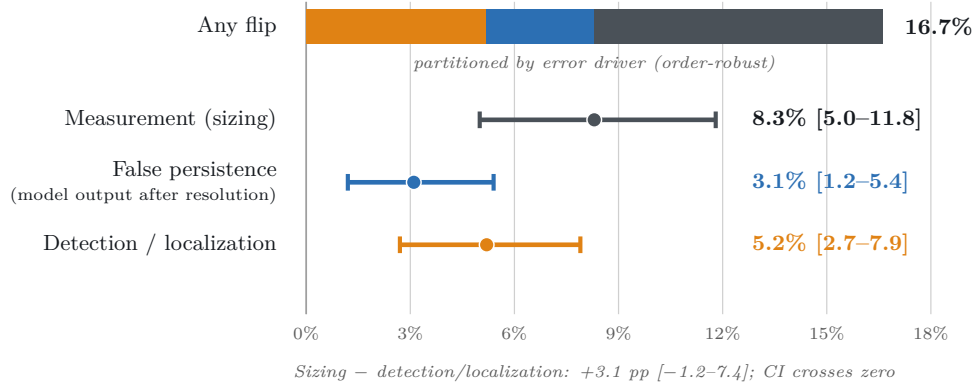


Fig. 7 Error-channel contributions to category disagreement. Shapley attribution divides the 16.7% disagreement rate among detection/localization (omission plus mislocalization), false persistence, and sizing error. False persistence is model output after reference resolution. Points show cohort rates with 95% patient-bootstrap confidence intervals. Sizing has the largest estimate but is not statistically separated from detection/localization combined. CI denotes confidence interval; pp, percentage points.

296 3.3 Exploratory direction of category disagreements

297 In the post-hoc direction analysis, 28 of 43 disagreements moved toward apparent
 298 benefit (a response that appears more favourable than the true underlying change)
 299 and 15 toward apparent harm (two-sided exact binomial $p = 0.066$). Omission tends
 300 to reduce measured burden, whereas false persistence tends to increase it; sizing errors
 301 and RECIST thresholds can move the category in either direction.

302 The counts changed to 32 versus 12 with one target per type and 26 versus 18 with
 303 five per type. The apparent imbalance was therefore not stable to target selection.
 304 This analysis was post hoc, was not adjusted for multiple analyses, and measured
 305 category movement rather than objective-response rate. It is descriptive and does not
 306 show that the model inflates objective-response rate [12].

3.4 Synthetic sizing plausibility control

Among 483 spatially matched follow-up units, the median held-out-fold diameter error was -1.9% (interquartile range -8.3 to $+2.9\%$; 5th–95th percentiles -29.4 to $+21.2\%$). Resampling these errors and applying them to expert follow-up diameters produced a mean sizing-only flip rate of 9.3% (central 95% simulation interval 6.2 – 12.4). This interval encompassed the observed 8.3% sizing contribution (Fig. 8).

Moving every follow-up boundary by -1.0 , -0.3 , 0 , $+0.3$, or $+1.0$ mm gave flip rates of 8.5% , 2.3% , 0% , 4.3% , and 14.0% , respectively. These shifts change diameter by -2.0 , -0.6 , 0 , $+0.6$, and $+2.0$ mm and do not depend on the model’s error distribution. Changing follow-up diameters by -10% or $+10\%$ gave flip rates of 7.8% and 12.8% . Applying the same change at baseline and follow-up gave only 0.8% because the error largely cancels in the ratio. These simulations show that errors of the observed size can plausibly produce a similar flip burden. They do not independently validate the attribution or establish causation.

3.5 A representative coalesced-conglomerate omission

Fig. 9 illustrates an omitted follow-up conglomerate that changed the target-lesion category from reference PD to model PR. This patient contributes to the reference-PD/model-PR cell ($n = 3$) in Fig. 6.

3.6 Aggregate sensitivity analyses

Across nine PR/PD threshold pairs, disagreement ranged from 15.5% to 17.8% ; changing nodal measurement from the caliper to a PCA minor-axis estimate changed it from 16.7% to 15.9% (Fig. 10).

Changing target selection produced similar cohort rates: one, two (primary), or five targets per type gave 17.1% , 16.7% , and 17.1% , respectively. Nevertheless, disagreement status changed for 27 patients under the one-per-type rule and 15 under

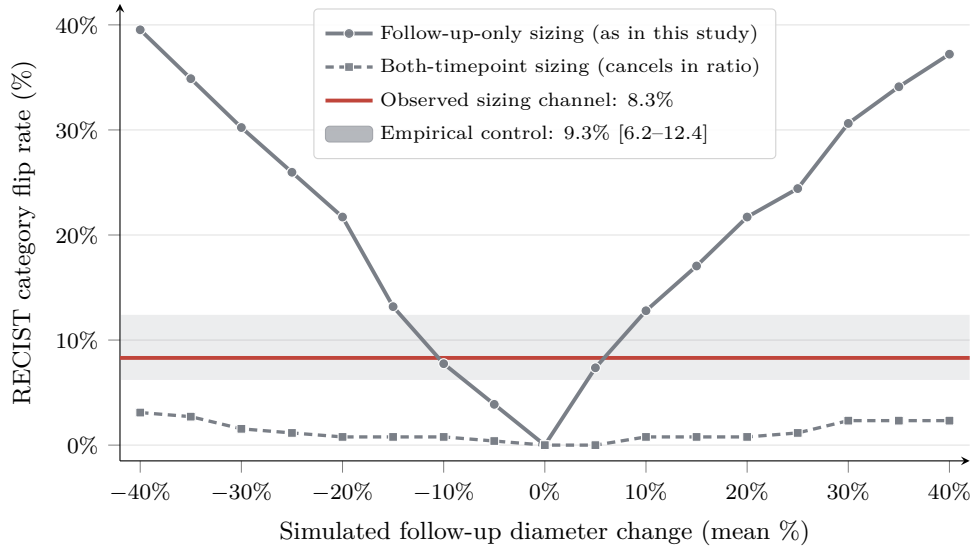


Fig. 8 Sizing plausibility control. Follow-up-only diameter changes (grey solid) show how sizing error affects the category. Resampling observed errors from 483 spatially matched units gave a mean flip rate of 9.3% (central 95% simulation interval 6.2–12.4; shaded band), which encompassed the observed 8.3% sizing contribution (red line). Applying the same factor at both timepoints (grey dashed) largely cancels.

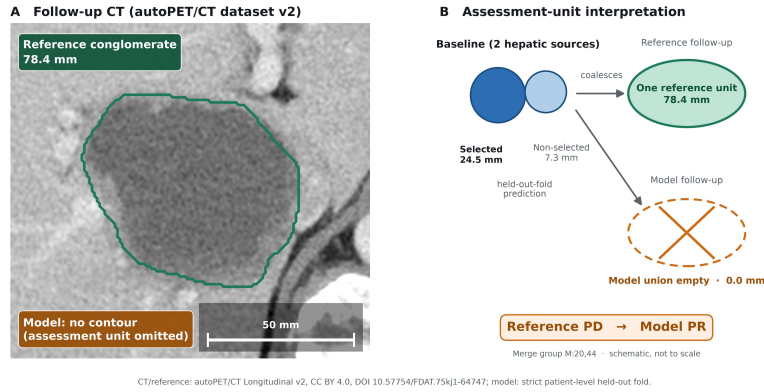


Fig. 9 Prompted omission of a coalesced liver target. This is the fourth error state; the other three are shown in Fig. 4. A 24.5-mm selected target coalesced with a 7.3-mm non-selected lesion (below the 10-mm measurability threshold) into a 78.4-mm reference conglomerate; only the selected source contributed to baseline SLD. Despite registration-propagated prompts for both sources, LongiSeg produced no merge-group voxels, changing reference progressive disease to model partial response. The white bar in panel A is a 50 mm scale bar. The image comes from autoPET/CT longitudinal dataset version 2 (CC BY 4.0; DOI: 10.57754/FDAT.75kj1-64747) [15]; official files were CRC- and geometry-checked, cropped, and annotated by the authors. SLD denotes sum of diameters.

the five-per-type rule. The deliberately inclusive skeletal upper bound gave 18.6% (Table 3). The skeletal row is an upper bound because the dataset does not identify which skeletal lesions have a RECIST-measurable soft-tissue component.

Table 3 Target-selection sensitivity on the fixed 258-patient cohort. Benefit:harm is the post-hoc count of disagreements moving toward apparent benefit versus apparent harm.

Selection rule	Disagreements	Rate (%)	Sizing (%)	Benefit:harm
One per lesion type, five total	44/258	17.1	9.5	32:12
Primary: two per type, five total	43/258	16.7	8.3	28:15
Five per type, five total	44/258	17.1	7.5	26:18
Skeletal upper bound	48/258	18.6	9.7	—

All 43 flips occurred among the 256 patients whose selected sources used only propagated prompts, a rate of 43/256 (16.8%); neither of the two fallback-affected patients flipped. Tightening the spatial-match rule preserved the 16.7% disagreement rate but shifted 0.4 percentage points from sizing to detection/localization (Fig. 10). Fold-specific rates ranged from 9.8% to 23.5%, whereas leave-one-fold-out rates ranged from 15.0% to 18.4%; no single fold drove the result.

4 Discussion

We evaluated a narrow assisted task. LongiSeg was given the expert baseline lesion and a follow-up location prompt, then asked to draw the follow-up lesion. Even with this support, its target-lesion category disagreed with the expert-mask reference in 16.7% of patients, *about one in six*. Because the baseline measurement was always expert-derived, these disagreements isolate follow-up tracking and measurement errors. They describe one prompt-assisted model, not longitudinal segmentation models as a class.

Previous automated-RECIST studies have reported overall agreement and reliability [6]. Our study asks a different question: which segmentation errors produce those disagreements?

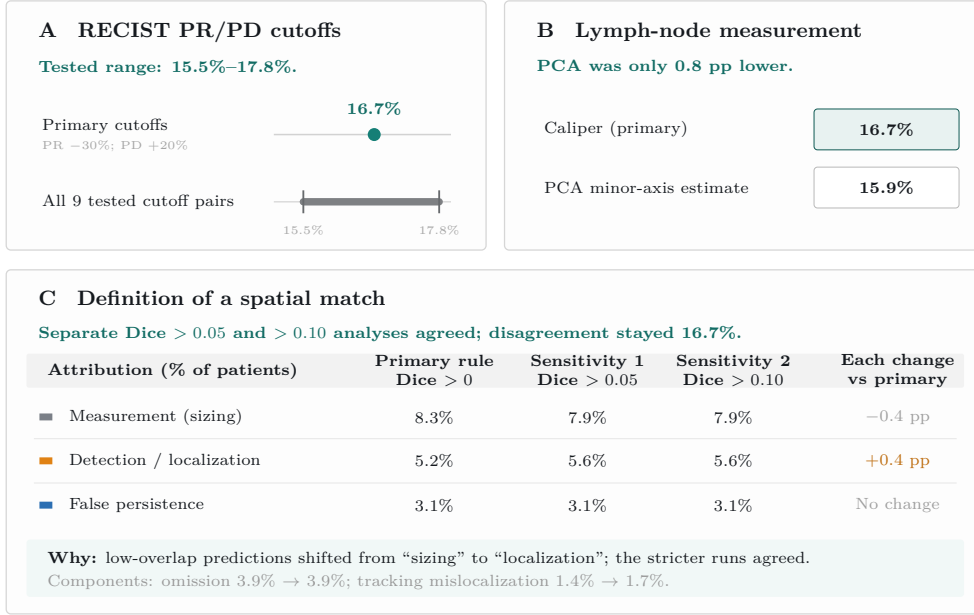


Fig. 10 Sensitivity analyses. Each panel compares the primary analysis with reasonable alternatives. **(A)** Across nine RECIST threshold pairs, category disagreement ranged from 15.5% to 17.8%; the primary thresholds gave 16.7%. **(B)** Changing nodal measurement from the caliper to the PCA estimate changed disagreement from 16.7% to 15.9%. **(C)** Tightening the spatial-match rule did not change which patients disagreed or the 16.7% disagreement rate; it only reassigned a small contribution from sizing to detection/localization. Dice > 0.05 and > 0.10 produced identical attribution estimates. Values are percentages of the 258-patient cohort; pp denotes percentage points.

4.1 Measurement precision matters alongside localization

Missed disease is the usual safety concern in automated response assessment. Here, however, sizing error in lesions that the model located had the largest estimate. It exceeded persistence and omission separately but was not statistically separated from detection/localization combined (+3.1 percentage points, 95% CI −1.2 to 7.4). Sizing was therefore important, but not unequivocally dominant. The data also do not support treating detection as the only relevant failure.

The sizing simulations produced a compatible number of flips when diameter errors matched those observed in the model. This agrees with the literature on human measurement variability, which shows that RECIST depends on how lesions are measured [3, 8, 10, 23]. Our results extend that concern to automated segmentation:

measurement error can change the category even when the correct lesion is found. Moving every follow-up boundary inward by 1 mm or outward by 1 mm, equivalent to diameter changes of -2 and $+2$ mm, produced flip rates of 8.5% and 14.0%. Automated response systems should therefore audit boundary accuracy as well as lesion coverage. Volumetric measures, which do not depend on one diameter, are one possible alternative [24–26].

4.2 The direction finding is exploratory

More disagreements moved toward apparent benefit than apparent harm (28 versus 15), but this post-hoc pattern changed with the target-selection rule. It is descriptive, not a confirmatory test of bias. We did not evaluate objective-response rate, and movement between target-lesion categories cannot establish objective-response-rate inflation. The model also changed individual categories in both directions.

4.3 Limitations

The endpoint covers only target lesions; it is not overall RECIST response. We did not assess non-target disease or autonomous new-lesion detection. New lesions alone can define progression [27], but we used expert new-lesion status only as a descriptive subgroup. The comparator was an expert mask, not an independent radiologist read.

The model received favourable prompt support: an expert baseline mask and a challenge-supplied follow-up location. Of 683 selected sources, 678 locations were registration-propagated and five were follow-up-centroid fallbacks. Propagated locations can miss lesions that move or change size [19]. The findings therefore concern prompted tracking and measurement, not autonomous detection; the fallback comparison was descriptive and unpowered.

The cohort came from one centre and one disease, metastatic melanoma, so external replication is needed. Each patient’s checkpoint excluded that patient during weight fitting, but the LongiSeg configuration had already been selected by cross-validation

on the same development cohort [16]. This remains an internal evaluation and may retain configuration-selection optimism. Our decomposition also reflects a single segmentation architecture. Segmentation accuracy in CT depends on the network that is chosen; comparisons of 2D and 3D U-Net configurations for abdominal CT report Dice differences of several percentage points [28, 29]. The relative contribution of sizing, detection, and persistence errors may therefore differ for other backbones.

Targets were selected by an algorithm rather than expert target flags, and lesion type only approximated organ. Although the overall disagreement rate changed little, individual status changed in 27 patients with one target per type and 15 with five. The nodal short axis was also a mask-based caliper approximation of a reader’s measurement; the PCA alternative changed the flip rate from 16.7% to 15.9%. For bone lesions, the dataset could not identify the lytic or mixed lesions with a measurable soft-tissue component that RECIST permits. We excluded all bone-typed lesions in the main analysis. The deliberately over-inclusive skeletal upper bound gave 18.6%, but it cannot replace proper bone-lesion phenotyping.

The analysis required recovery of initially missed volumes and downloads. One patient remained technically non-evaluable because six propagated prompts fell outside a resampled image volume. Counting this patient as concordant or discordant would give 43/259 (16.6%) or 44/259 (17.0%), respectively; neither bound changes the interpretation.

Finally, all analyses were post hoc. Category disagreements in this target-lesion measure do not prove changes in clinical decisions or patient outcomes.

4.4 Implications and future work

Evaluations of automated response assessment should report why categories disagree, not only how often. They should audit measurement precision as well as lesion coverage. Future work should use a powered, multi-interval design to test whether tracking

worsens as the interval between scans grows. It should also compare with an independent radiologist read and assess overall response, including non-target disease and new lesions.

5 Conclusions

Longitudinal segmentation errors changed the mask-derived RECIST 1.1 target-lesion category in 16.7% of patients (Wilson 95% CI 12.6–21.7). Sizing had the largest estimate and exceeded persistence and omission separately, but it was not statistically separated from detection/localization combined. No channel clearly dominated. The post-hoc excess of changes toward apparent benefit depended on target selection, requires confirmation, and does not show objective-response-rate inflation.

These results show that measurement error in correctly located lesions matters alongside missed or mislocalized lesions. The sizing simulations supported the plausibility of a contribution at this scale but did not independently validate it. The findings are conditional on a correct expert baseline and describe one prompt-assisted model. They do not cover overall RECIST response, autonomous detection, or clinical impact.

List of abbreviations

AI: Artificial intelligence; CI: Confidence interval; CLAIM: Checklist for Artificial Intelligence in Medical Imaging; CR: Complete response; CT: Computed tomography; OOF: Out-of-fold; PCA: Principal component analysis; PD: Progressive disease; PR: Partial response; RECIST: Response Evaluation Criteria in Solid Tumours; SD: Stable disease; SLD: Sum of lesion diameters; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate: The source data collection and this retrospective secondary analysis were conducted in accordance with the Declaration of Helsinki. Source collection for the public, fully de-identified data was approved by the Ethics Committee of the University of Tübingen (project 700/2024BO2), which waived informed consent and permitted anonymized publication; no further approval or consent was required [13, 14].

Consent for publication: The fully de-identified CT excerpts in Figs. 1, 4 and 9 contain no identifiers; the source waiver permitted anonymized scientific publication, so no additional consent was required [14].

Availability of data and materials: We analysed Longitudinal-CT version 1 (CC BY-NC 4.0; DOI: 10.57754/FDAT.qwsry-7t837); version 2 (CC BY 4.0; DOI: 10.57754/FDAT.75kj1-64747) supplied the topology metadata and the de-identified CT excerpts in Figs. 1, 4 and 9 [13–15]. This article is self-contained: every algorithm is specified in pseudocode in Appendix A, the locked reported quantities are collected in Table A1, and every threshold, cohort count and audit needed to reproduce the reported percentages, intervals and counts is given in Methods and Results. No supplementary file is required. The identifier-free 258-row patient-level result table and the locked aggregate outputs are available from the corresponding author on reasonable request. Runnable analysis software is not distributed. The authors’ original material is licensed under Apache 2.0. No separate project homepage or archived-code identifier is claimed. LongiSeg source and checkpoints remain at their cited locations; identifiers, patient-level model masks, third-party source, and checkpoints are not redistributed [16, 30].

Competing interests: The authors declare that they have no competing interests.

Funding: This work was funded by the Deutsche Forschungsgemeinschaft (DFG), project MA 6791/1-1; the funder had no role in study design, analysis, interpretation, or manuscript preparation.

Authors’ contributions: S.A. conceived and implemented the analysis, performed the experiments, and drafted the manuscript; N.C. contributed methodology and revision; A.M. supervised the work, acquired funding, and revised the manuscript. All authors approved the final manuscript.

466 **Acknowledgements:** The authors thank the autoPET challenge organizers and data
 467 providers for making the longitudinal CT dataset public.

468 Appendix A Analysis specification

469 This appendix specifies, in pseudocode, every algorithm behind the reported numbers,
 470 so that the article is reproducible without a supplementary file. It is a specification, not
 471 an executable release: no runnable source code, dependency manifest or verification
 472 program is distributed. Reference measurements always come from the expert masks at
 473 both timepoints; model measurements use the expert baseline mask and the predicted
 474 follow-up mask. Nothing below uses the expert follow-up mask as a model input, which
 475 is only the evaluation reference.

476 A.1 Notation

477 For one patient: $\mathcal{T}$ is the set of selected baseline target lesions; L indexes a lesion label;
 478 $\text{node}(L)$ is true for a pathological lymph node; b_L and g_L are the reference baseline
 479 and reference follow-up diameters of L in mm; m_L is the model follow-up diameter;
 480 d_L is the Dice overlap between the model and reference voxels carrying label L ; and
 481 τ is the spatial-match gate ($\tau = 0$ in the primary analysis, so any shared voxel counts
 482 as a match). Thresholds are the RECIST 1.1 values: PR = -30%, PD = +20% with
 483 a +5 mm absolute floor, and a 10 mm nodal ceiling for complete response.

484 A.2 Target selection

Algorithm A1 Select target lesions from the expert baseline mask

Input: expert baseline mask, dataset lesion type per label, caps $k_{\text{type}} = 2$ and $k_{\text{total}} = 5$

Output: ordered target set $\mathcal{T}$

```

1: measure every baseline label: short axis if nodal, else longest axial diameter
2: discard labels below the measurability floor (nodal < 15 mm, non-nodal < 10 mm)
3: discard every bone-typed label  $\triangleright$  the release does not identify which bone lesions
   carry a RECIST-measurable soft-tissue component
4: sort the survivors by descending diameter
5:  $\mathcal{T} \leftarrow \emptyset$ ;  $c[\cdot] \leftarrow 0$ 
6: for each label  $L$  in sorted order do
7:   if  $|\mathcal{T}| < k_{\text{total}}$  and  $c[\text{type}(L)] < k_{\text{type}}$  then
8:      $\mathcal{T} \leftarrow \mathcal{T} \cup \{L\}$ ;  $c[\text{type}(L)] \leftarrow c[\text{type}(L)] + 1$ 
9:   end if
10: end for
```

485 All labels carrying the dataset's *unclear* type share one bucket for the per-type cap.
 486 A patient with $\mathcal{T} = \emptyset$ is not response-evaluable.

487 A.3 Assessment units and coalescence

Algorithm A2 Build follow-up assessment units

Input: $\mathcal{T}$, per-lesion metadata, follow-up volumes

Output: assessment units $\mathcal{U}$, each with a baseline sum and one reference measurement

```

1: group labels that share a follow-up image and centre of gravity      ▷ version 1
   encoding
2: cross-check every group against topology_class = MERGING and merged_into ▷
   version 2
3: for each merge group  $G$  with at least one selected source do
4:   create one unit  $u$  with sources  $G \cap \mathcal{T}$ 
5:   baseline sum of  $u \leftarrow \sum_{L \in G \cap \mathcal{T}} b_L$ 
6:   reference measurement of  $u \leftarrow$  measurement of the surviving reference label
7:   model measurement of  $u \leftarrow$  measurement of the union of predicted labels over
   all of  $G$ 
8: end for
9: for each selected label not in any merge group do
10:  create one unit with that label alone
11: end for

```

488 Taking the union over the full merge group, rather than the single surviving label,
489 prevents a label renaming from being scored as a disappearance. Where an examination
490 contains overlapping body-region volumes, each unit is measured in every volume and
491 its largest measurement is retained, so a lesion appearing only in a later volume is not
492 treated as absent.

493 A.4 Measurement and response category

Algorithm A3 Measure one mask component

Input: component voxels, voxel spacing, $\text{node}(L)$

```

1: if  $\text{node}(L)$  then
2:   take the largest-area axial slice; compute its convex hull
3:   find the longest in-slice caliper; return the greatest width perpendicular to it
   ▷ short axis
4: else
5:   return the longest in-slice diameter over all axial slices      ▷ longest diameter
6: end if

```

Algorithm A4 Target-lesion response category

Input: $\mathcal{T}$, node, baseline diameters b , follow-up diameters f

Output: CR, PR, SD or PD

```
1: if  $\mathcal{T} = \emptyset$  then
2:   return not evaluable
3: end if
4:  $S_b \leftarrow \sum_{L \in \mathcal{T}} b_L$ ;  $S_f \leftarrow \sum_{L \in \mathcal{T}} f_L$   $\triangleright$  a target absent from  $f$  counts as 0; nodes
   contribute short axis even below 10 mm
5: if  $f_L \leq 0$  for every non-nodal  $L$  and  $f_L < 10$  mm for every nodal  $L$  then
6:   return CR
7: end if
8: if  $S_b \leq 0$  then
9:   return SD
10: end if
11:  $\Delta \leftarrow (S_f - S_b)/S_b$ 
12: if  $\Delta \leq -0.30$  then
13:   return PR
14: end if
15: if  $\Delta \geq +0.20$  and  $S_f - S_b \geq 5$  mm then
16:   return PD
17: end if
18: return SD
```

494 Because the study compares baseline with the first follow-up only, the nadir coincides
495 with baseline. The reference category uses $f_L = g_L$; the model category uses $f_L = m_L$.
496 A disagreement is any patient whose two categories differ.

Algorithm A5 Overlap-gated error state per assessment unit

Input: g_L , m_L , d_L , gate τ **Output:** one state per unit

```

1: matched  $\leftarrow (d_L > \tau)$ 
2: if  $g_L > 0$  and  $m_L \leq 0$  then
3:   return prompted omission
4: else if  $g_L > 0$  and  $m_L > 0$  and not matched then
5:   return tracking mislocalization
6: else if  $g_L > 0$  and  $m_L > 0$  and matched and  $|m_L - g_L| > 0$  then
7:   return measurement (sizing) error
8: else if  $g_L \leq 0$  and  $m_L > 0$  then
9:   return false persistence
10: else
11:   return agreement
12: end if

```

498 Sizing is therefore defined only where the model and reference share voxels: a non-
499 overlapping prediction is a localization failure, never a measurement failure. The gate
500 changes only this label. The response category always uses the measurements from the
501 model's actual mask, so the disagreement rate cannot move with τ by construction.

A.6 Order-robust attribution

Algorithm A6 Shapley attribution of a category disagreement

Input: unit states, reference category c_{ref} , channel set C of size 4

Output: contribution $\phi_c \geq 0$ for each channel, with $\sum_c \phi_c = 1$

```

1: function CORRECT( $S$ )
2:    $f \leftarrow m$ 
3:   for each unit whose state  $c$  lies in  $S$  do
4:     if  $c$  is false persistence then
5:       remove the unit from  $f$  ▷ drop the false positive
6:     else
7:        $f_L \leftarrow g_L$  ▷ omission, mislocalization and sizing all restore the reference
       diameter
8:     end if
9:   end for
10:  return  $\mathbb{K}[\text{CATEGORY}(f) \neq c_{\text{ref}}]$ 
11: end function
12: for each of the  $2^4 = 16$  subsets  $S \subseteq C$  do
13:    $v(S) \leftarrow \text{CORRECT}(S)$ 
14: end for
15:  $\phi_c \leftarrow 0$  for all  $c$ 
16: for each of the  $4! = 24$  orderings  $\pi$  of  $C$  do
17:    $S \leftarrow \emptyset$ 
18:   for each channel  $c$  in  $\pi$  do
19:      $\phi_c \leftarrow \phi_c + (v(S) - v(S \cup \{c\})); \quad S \leftarrow S \cup \{c\}$ 
20:   end for
21: end for
22:  $\phi_c \leftarrow \phi_c / 24$  for all  $c$ 

```

Channels absent from a patient are null players and receive zero. For all 43 disagreements v was monotone non-increasing and $v(C) = 0$, so contributions were non-negative and summed to one per patient. Cohort percentages are the summed patient contributions divided by 258. Omission and mislocalization are additionally reported together as detection/localization. [Table 1](#) works one disagreement through by hand.

509 A.7 Sizing plausibility simulations

Algorithm A7 Follow-up-only sizing simulations

Input: reference baseline and follow-up diameters; observed relative errors from matched units

- 1: **multiplicative:** for $\alpha \in \{0.60, 0.65, \dots, 1.40\}$, set $f_L \leftarrow \alpha g_L$ for present units and recompute the category
- 2: **boundary:** for $\delta \in \{0, \pm 0.15, \pm 0.30, \pm 0.60, \pm 1.0, \pm 1.5\}$ mm, set $f_L \leftarrow g_L + 2\delta$ $\triangleright$ a boundary shift moves the diameter twice
- 3: **empirical:** repeat 2000 times with seed 20260711
- 4: for each present unit, draw a relative error with replacement from the matched-unit pool and apply it to g_L
- 5: recompute every category and record the flip rate
- 6: report the mean and the 2.5th and 97.5th percentiles of the 2000 rates
- 7: **common mode:** apply the same α at baseline *and* follow-up to show cancellation in the ratio

510 Baseline diameters stay at the expert reference and resolved targets stay absent,
 511 matching the main analysis. Errors are drawn independently per unit, so the empir-
 512 ical simulation does not model within-patient correlation and is model-specific; the
 513 multiplicative and boundary simulations are model-agnostic.

514 A.8 Sensitivity sweeps

Algorithm A8 Sensitivity sweeps on the fixed 258-patient cohort

- 1: extract every measurement once per patient, with both nodal estimators
- 2: **thresholds:** recompute over nine PR/PD threshold pairs
- 3: **nodal estimator:** caliper short axis versus a principal-component minor-axis estimate
- 4: **spatial gate:** $\tau \in \{0, 0.05, 0.10\}$ $\triangleright$ affects attribution only
- 5: **target selection:** $k_{\text{type}} \in \{1, 2, 5\}$ with $k_{\text{total}} = 5$ held fixed
- 6: **skeletal upper bound:** treat every size-eligible bone mask as measurable
- 7: **prompt provenance:** recompute after excluding patients with any follow-up-centroid fallback prompt
- 8: **folds:** report per-fold and leave-one-fold-out rates

515 The prompt-provenance audit compared every challenge click against both candidate
 516 coordinates: of 4298 follow-up clicks across the cohort, 3986 matched the propa-
 517 gated coordinate and 312 matched the follow-up-annotated coordinate. Among the
 518 683 selected sources, 678 were propagated and five used the fallback.

519 A.9 Locked results

520 Table A1 is the locked set of reported quantities, so that every percentage, interval
521 and count in the article can be checked in one place.

Table A1 Locked primary result and channel contributions ($n = 258$).

Quantity	Value	95% interval
Category disagreement	43/258 (16.7%)	12.6–21.7 (Wilson)
Agreement; Cohen κ	215/258 (83.3%); 0.757	—
Measurement (sizing)	8.3%	5.0–11.8
Detection/localization	5.2%	2.7–7.9
prompted omission	3.9%	1.7–6.4
tracking mislocalization	1.4%	0.2–2.9
False persistence	3.1%	1.2–5.4
Sizing – detection/localization	+3.1 pp	–1.2 to 7.4
Sizing – false persistence	+5.2 pp	1.4–9.3
Sizing – prompted omission	+4.5 pp	0.4–8.5
False persistence – detection/localization	–2.1 pp	–5.6 to 1.2
Direction (benefit : harm)	28 : 15	exact binomial $p = 0.066$
Empirical sizing simulation	9.3% mean	6.2–12.4 (simulation)

522 References

- 523 [1] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al.
524 New response evaluation criteria in solid tumours: revised RECIST guideline
525 (version 1.1). European Journal of Cancer. 2009;45(2):228–247. <https://doi.org/10.1016/j.ejca.2008.10.026>.
526
- 527 [2] U S Food and Drug Administration. Clinical Trial Endpoints for
528 the Approval of Cancer Drugs and Biologics: Guidance for Indus-
529 try. Silver Spring, MD: U.S. Department of Health and Human
530 Services, Food and Drug Administration; 2018. Available from:
531 [https://www.fda.gov/regulatory-information/search-fda-guidance-documents/](https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics)
532 [clinical-trial-endpoints-approval-cancer-drugs-and-biologics](https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics).

- 533 [3] Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS,
534 et al. Interobserver and Intraobserver Variability in Measurement of Non-Small-
535 Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor Response.
536 Journal of Clinical Oncology. 2003;21(13):2574–2582. [https://doi.org/10.1200/
537 JCO.2003.01.144](https://doi.org/10.1200/JCO.2003.01.144).
- 538 [4] Yoon SH, Kim KW, Goo JM, Kim DW, Hahn S. Observer variability in RECIST-
539 based tumour burden measurements: a meta-analysis. European Journal of
540 Cancer. 2016;53:5–15. <https://doi.org/10.1016/j.ejca.2015.10.014>.
- 541 [5] Leszczyńska A, Seweryn M, Obuchowicz R, Strzelecki M, Piórkowski A,
542 Potocki PM. Artificial Intelligence for RECIST-Based Radiologic Treatment
543 Response Assessment in Solid Tumors: A Systematic Review of Imaging- and
544 Report-Derived Approaches. Cancers. 2026;18(5):808. [https://doi.org/10.3390/
545 cancers18050808](https://doi.org/10.3390/cancers18050808).
- 546 [6] Dahm IC, Kolb M, Altmann S, Nikolaou K, Gatidis S, Othman AE, et al. Reliabil-
547 ity of Automated RECIST 1.1 and Volumetric RECIST Target Lesion Response
548 Evaluation in Follow-Up CT—A Multi-Center, Multi-Observer Reading Study.
549 Cancers. 2024;16(23):4009. <https://doi.org/10.3390/cancers16234009>.
- 550 [7] Xia Y, Zhou J, Xun X, Johnston L, Wei T, Gao R, et al. Deep learning
551 for oncologic treatment outcomes and endpoints evaluation from CT scans in
552 liver cancer. npj Precision Oncology. 2024;8(1):263. [https://doi.org/10.1038/
553 s41698-024-00754-z](https://doi.org/10.1038/s41698-024-00754-z).
- 554 [8] Cornelis FH, Martin M, Saut O, Buy X, Kind M, Palussiere J, et al. Precision
555 of manual two-dimensional segmentations of lung and liver metastases and its
556 impact on tumour response assessment using RECIST 1.1. European Radiology
557 Experimental. 2017;1(1):16. <https://doi.org/10.1186/s41747-017-0015-4>.

- 558 [9] Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, et al. Vari-
559 ability of Lung Tumor Measurements on Repeat Computed Tomography Scans
560 Taken Within 15 Minutes. *Journal of Clinical Oncology*. 2011;29(23):3114–3119.
561 <https://doi.org/10.1200/JCO.2010.33.7071>.
- 562 [10] McErlean A, Panicek DM, Zabor EC, Moskowitz CS, Bitar R, Motzer RJ, et al.
563 Intra- and Interobserver Variability in CT Measurements in Oncology. *Radiology*.
564 2013;269(2):451–459. <https://doi.org/10.1148/radiol.13122665>.
- 565 [11] Zhou J, Xia Y, Xun X, Yu Z. Deep Learning-Based Detect-Then-Track Pipeline
566 for Treatment Outcome Assessments in Immunotherapy-Treated Liver Cancer.
567 *Journal of Imaging Informatics in Medicine*. 2025;38(1):380–393. [https://doi.org/](https://doi.org/10.1007/s10278-024-01132-8)
568 [10.1007/s10278-024-01132-8](https://doi.org/10.1007/s10278-024-01132-8).
- 569 [12] Dello Russo C, Cappoli N, Pilunni D, Navarra P. Local Investigators Significantly
570 Overestimate Overall Response Rates Compared to Blinded Independent Central
571 Reviews in Phase 2 Oncology Trials. *The Journal of Clinical Pharmacology*.
572 2021;61(6):810–819. <https://doi.org/10.1002/jcph.1790>.
- 573 [13] Küstner T, Peisen F, Gatidis S, Wagner A, Megne O, Othman A, et al.
574 Longitudinal-CT (autoPET/CT IV Task 2 dataset). University of Tübingen,
575 FDAT Research Data Repository. Longitudinal whole-body CT (melanoma),
576 300 patients, baseline and follow-up timepoints with expert radiologist lesion
577 masks. License: CC BY-NC 4.0. Available from: [https://doi.org/10.57754/FDAT.](https://doi.org/10.57754/FDAT.qwsry-7t837)
578 [qwsry-7t837](https://doi.org/10.57754/FDAT.qwsry-7t837).
- 579 [14] Gatidis S, Peisen F, Wagner A, Megne Choudja PO, Othman A, Sanner A, et al.
580 A longitudinal whole-body CT dataset with manually annotated tumor lesions.
581 *Scientific Data*. 2026;13. <https://doi.org/10.1038/s41597-026-07466-y>.

- [15] Küstner T, Peisen F, Gatidis S, Wagner A, Megne O, Othman A, et al. Longitudinal-CT (autoPET/CT IV Task 2 dataset), version 2. University of Tübingen, FDAT Research Data Repository. Version 2 with explicit lesion-topology metadata; license: CC BY 4.0. Available from: <https://doi.org/10.57754/FDAT.75kj1-64747>.
- [16] Kirchhoff Y, Rokuss M, Isensee F, Maier-Hein KH. Promptable Longitudinal Lesion Segmentation in Whole-Body CT. AutoPET/CT IV Challenge (Task 2), MICCAI 2025; code: <https://github.com/MIC-DKFZ/LongiSeg/tree/autoPET>; evaluated from the autoPET branch at commit ee5296ee4b0ebc4f56aa916c34215edb9629dace, with the local inference edge-clamping patch supplied in the reproducibility materials. Available from: <https://arxiv.org/abs/2509.00613>.
- [17] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021;18(2):203–211. <https://doi.org/10.1038/s41592-020-01008-z>.
- [18] Cai J, Tang Y, Yan K, Harrison AP, Xiao J, Lin G, et al. Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 15159–15169. Doi: 10.1109/CVPR46437.2021.01491.
- [19] Rocholl N, Smit E, Prokop M, Hering A. Unstable Prompts, Unreliable Segmentations: A Challenge for Longitudinal Lesion Analysis. In: Learning with Longitudinal Medical Images and Data. vol. 16184 of Lecture Notes in Computer Science. Cham: Springer; 2026. p. 13–23. Doi: 10.1007/978-3-032-16128-4_2; arXiv:2507.19230.

- 607 [20] Shapley LS. A value for n -person games. In: Kuhn HW, Tucker AW, editors. Con-
608 tributions to the Theory of Games, Volume II. No. 28 in Annals of Mathematics
609 Studies. Princeton, NJ: Princeton University Press; 1953. p. 307–317.
- 610 [21] Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E,
611 et al. Metrics reloaded: recommendations for image analysis validation. *Nature*
612 *Methods*. 2024;21(2):195–212. <https://doi.org/10.1038/s41592-023-02151-z>.
- 613 [22] Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D,
614 Kavur AE, et al. Understanding metric-related pitfalls in image analysis
615 validation. *Nature Methods*. 2024;21(2):182–194. <https://doi.org/10.1038/s41592-023-02150-0>.
616
- 617 [23] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al.
618 Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans
619 of Patients with Non–Small Cell Lung Cancer. *Radiology*. 2009;252(1):263–272.
620 <https://doi.org/10.1148/radiol.2522081593>.
- 621 [24] Hayes SA, Pietanza MC, O’Driscoll D, Zheng J, Moskowitz CS, Kris MG, et al.
622 Comparison of CT volumetric measurement with RECIST response in patients
623 with lung cancer. *European Journal of Radiology*. 2016;85(3):524–533. <https://doi.org/10.1016/j.ejrad.2015.12.019>.
624
- 625 [25] Lubner MG, Stabo N, Lubner SJ, Munoz Del Rio A, Song C, Pickhardt PJ.
626 Volumetric versus unidimensional measures of metastatic colorectal cancer in
627 assessing disease response. *Clinical Colorectal Cancer*. 2017;16(4):324–333. <https://doi.org/10.1016/j.clcc.2017.03.009>.
628
- 629 [26] Zhao B, Obuchowski N, Yang H, Chou Y, Ma H, Guo P, et al. Compar-
630 ing quantitative imaging biomarker alliance volumetric CT classifications with

- 631 RECIST response categories. *Radiology Advances*. 2025;2(1):umaf001. <https://doi.org/10.1093/radadv/umaf001>.
632
- 633 [27] Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al.
634 iRECIST: guidelines for response criteria for use in trials testing immunothe-
635 rapies. *The Lancet Oncology*. 2017;18(3):e143–e152. [https://doi.org/10.1016/](https://doi.org/10.1016/S1470-2045(17)30074-8)
636 [S1470-2045\(17\)30074-8](https://doi.org/10.1016/S1470-2045(17)30074-8).
- 637 [28] Zettler N, Mastmeyer A. Comparison of 2D vs. 3D U-Net organ segmentation
638 in abdominal 3D CT images. In: *Proceedings of the International Confer-*
639 *ence on Computer Graphics, Visualization and Computer Vision (WSCG)*; 2021.
640 *ArXiv:2107.04062*.
- 641 [29] Kern D, Klauck U, Ropinski T, Mastmeyer A. 2D vs. 3D U-Net abdominal organ
642 segmentation in CT data using organ bounds. In: *Medical Imaging 2021: Imaging*
643 *Informatics for Healthcare, Research, and Applications*. vol. 11601 of *Proceedings*
644 *of SPIE*. SPIE; 2021. p. 1160114.
- 645 [30] Kirchhoff Y. autoPET/CT IV: LongiLesionLocator. Zenodo. Available from:
646 <https://doi.org/10.5281/zenodo.17027189>.