Appendix A. Additional methodological details

Appendix A.1. Implementation and optimization of comparison baselines

For the CNN- and VAE-based regressors, image-segmentation pairs and segmentations alone were used as input, respectively, with the segmentations generated by nnU-Net (Sect. 5.4 of the main article). During preprocessing, inputs were first cropped around the foreground region of each segmentation, followed by resizing to a fixed input size required by the CNN, using padding and resampling where necessary. Random data augmentation included scaling, adding Gaussian noise, intensity rescaling, and mirroring. In addition, affine transformations were applied to the segmentation masks for the CNN-based regressor.

The CNN-based regressor was implemented using a U-Net encoder, followed by global average pooling and a linear layer to predict Dice accuracy. The model was trained using the AdamW optimizer with cosine-annealed weight decay to minimize the L2 loss. Input crop size, resampling, and batch size were optimized as hyper-parameters.

The VAE-based regressor was implemented as a convolutional encoder–decoder without skip connections, comprising five downsampling and upsampling stages and a latent space of dimensionality 256. During training, ground-truth masks were used as input, and an evidence lower bound (ELBO) loss was optimized combining a binary cross-entropy mask reconstruction loss and a Kullback-Leibler divergence term against a standard normal prior, weighted by $\beta = 0.001$. At test time, the ELBO loss was used as an uncertainty score, with higher values indicating poorer reconstruction and thus lower predicted segmentation quality. Additionally, a linear model was trained to map ELBO-loss values to predicted Dice scores.

The RF approach utilized nnU-Net-generated segmentations (Sect. 5.4 of the main article) together with their corresponding entropy maps to perform Dice accuracy regression. We compared a Radiomics- and a heuristic-based variant of the approach.

The Radiomics-based variant relied on features extracted from regions affected by segmentation errors using PyRadiomics. These regions were localized by applying an uncertainty threshold selected to maximize overlap with true error regions. The threshold was optimized on the validation split of the training set (Jungo et al., 2020).

The heuristic-based variant incorporated assumptions about the relevance of ROI size, morphology, and uncertainty distributions across foreground, background and boundary regions. Specifically, entropy values were averaged within these regions, and combined with foreground size and connected component count as input features for RF training. Including uncertainty distributions beyond the foreground region may provide additional information for segmentation error estimation. The heuristic-based variant proved more effective and was therefore employed in all experiments. In all cases, ensembles of 100 trees were trained using mean-squared error as the splitting criterion and bagging with bootstrap samples of the same size as the training set, drawn with replacement. No pruning was applied.

Appendix B. Interpretation of evaluation metrics

Appendix B.1. Conceptual differences between AURC and $AUPRC_F$

Segmentation selection and failure sampling capture distinct aspects of error-estimator performance. An estimator that performs better at selecting accurate segmentations in terms of AURC does not necessarily perform better at sampling failure cases in terms of AUPRC_F. For example, two error estimators may assign uncertainties to the same number of truly erroneous (1) and error-free (0) segmentations. Let $(0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0)$ and $(0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0)$ denote the segmentations arranged from low to high uncertainties, as produced by both error estimators. Estimator Two is overall better at selecting segmentations with low error risk in terms of the AURC ($0.26 < 0.30$, with lower values indicating better performance). However, it performs less effectively at sampling error cases without incorrectly presenting error-free cases, as measured by the $\mathrm{AUPRC}_\mathrm{F}$ ($0.53 < 0.61$, with higher values indicating better performance).

Appendix C. Additional results

Appendix C.1. Bootstrap comparisons of error estimator performance

Bootstrap experiments were conducted to extend the error estimation analyses detailed in Sect. 6 of the main article. Specifically, paired performance differences in alignment between error estimates and delineation accuracy, segmentation selection, and failure case sampling were determined between the best-performing PWD-based and non-PWD-based estimators. Bootstrap samples ($n = 10{,}000$) were drawn at the participant level, including only segmentation-uncertainty pairs associated with participants selected in each bootstrap sample, thereby accounting for the many-to-one mapping between segmentation-uncertainty pairs and participants. For each bootstrap sample, the paired performance difference between the two estimators was determined, and the corresponding 95% confidence interval (CI) was derived from the resulting bootstrap distribution. Beyond the paired comparisons, bootstrap performance estimates and CIs were also determined for each error estimator individually.

For the AURC measures, the reference frame depends on the characteristics of each bootstrap sample, namely the corresponding oracle and random AURC performance. To enable comparison across bootstrap samples, we therefore used excess AURC (E-AURC, Geifman and El-Yaniv, 2017), which measures the difference between the AURC of an error estimator and that of the oracle estimator within a given bootstrap sample.

The resulting paired performance differences and individual estimator performances are shown in Fig. C.1 and Tab. C.1,

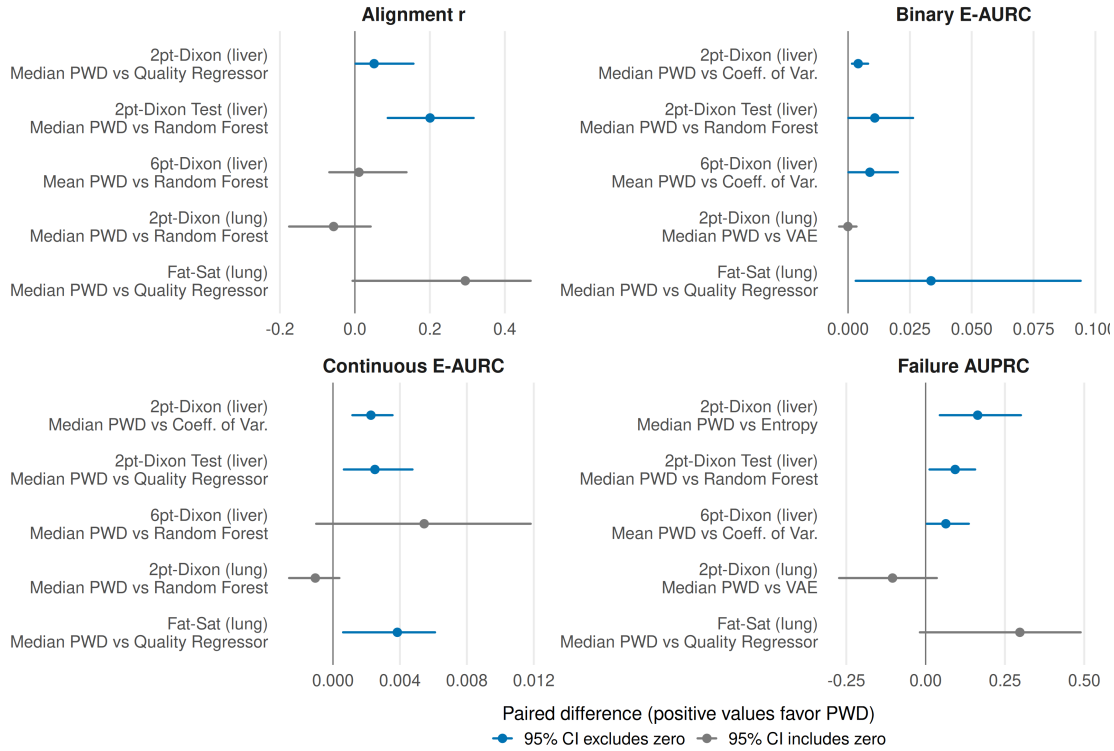


Figure C.1: **Bootstrap comparison of error estimator performance:** Paired bootstrap differences in alignment between error estimates and delineation accuracy, binary and continuous E-AURC, and AUPRC$_F$ between estimators support the use of pairwise-Dice (PWD)-based estimators for large-scale segmentation quality control. Across participant-level bootstrap samples (n = 10,000), PWD-based estimators were compared with the best-performing non-PWD estimator in each setting. The 95% confidence intervals (CIs) indicated a positive performance difference in favor of PWD-based estimators in 15 of 20 comparisons, with CIs excluding zero in 12 cases. Non-PWD estimators were favored in 3 of 20 comparisons (error-estimate and accuracy alignment, continuous E-AURC, and AUPRC$_F$). In none of these cases did the 95% CI exclude zero.

respectively. Across all evaluated performance measures and datasets, paired differences favored PWD-based estimators in 15 of 20 comparisons, with 95% CIs excluding zero in 12 cases. Non-PWD estimators were favored in 3 of 20 comparisons (error-estimate and accuracy alignment, continuous E-AURC, and AUPRC$_F$); in none of these cases did the 95% CI exclude zero. The remaining two comparisons showed comparable performance. Overall, these results further support PWD-based estimators for large-scale segmentation quality control.

For error-estimate and accuracy alignment, paired differences in Pearson r favored PWD-based estimators in three of five datasets (NAKO_EXP$_{\text{2pd_liver}}$, NAKO_EXP$_{\text{2pd_liver_test}}$, and COSYCONET_EXP$_{\text{test}}$), with 95% CIs excluding zero in the two 2pt-Dixon liver datasets (Fig. C.1, top-left plot). The non-PWD estimator was favored on NAKO_EXP$_{\text{2pd_lung}}$, although its CI included zero, while performance was comparable on NAKO_EXP$_{\text{6pd_liver}}$.

For threshold-dependent segmentation selection, binary E-AURC favored PWD-based estimators in four of five datasets (NAKO_EXP$_{\text{2pd_liver}}$, NAKO_EXP$_{\text{2pd_liver_test}}$, NAKO_EXP$_{\text{6pd_liver}}$, and COSYCONET_EXP$_{\text{test}}$), with 95% CIs excluding zero in all four cases (Fig. C.1, top-right plot). Performance was comparable on NAKO_EXP$_{\text{2pd_lung}}$.

For continuous segmentation selection, E-AURC favored PWD-based estimators in four of five datasets, with 95% CIs excluding zero for NAKO_EXP$_{\text{2pd_liver}}$, NAKO_EXP$_{\text{2pd_liver_test}}$, and COSYCONET_EXP$_{\text{test}}$ (Fig. C.1, bottom-left plot). On NAKO_EXP$_{\text{6pd_liver}}$, the paired difference also favored PWD-based estimators, although the CI included zero. The non-PWD estimator was favored on NAKO_EXP$_{\text{2pd_lung}}$, likewise with a CI including zero.

For failure case sampling, paired differences in AUPRC$_F$ favored PWD-based estimators in four of five datasets, with 95% CIs excluding zero for all three NAKO liver datasets (NAKO_EXP$_{\text{2pd_liver}}$, NAKO_EXP$_{\text{2pd_liver_test}}$, and NAKO_EXP$_{\text{6pd_liver}}$; Fig. C.1, bottom-right plot). The non-PWD estimator was favored on NAKO_EXP$_{\text{2pd_lung}}$, although its CI included zero, while PWD-based estimators were favored on COSYCONET_EXP$_{\text{test}}$ with a CI likewise including zero.

Appendix C.2. Summary statistics for cohort-level segmentation quality-control analyses in NAKO and COSYCONET

Bootstrap sampling and significance tests were conducted for the reported uncertainty distributions, organ volume and fat measurements, and lung mask characteristics (Sects. 6.5 and 6.6 of the main article). Bootstrap samples (n = 10,000) were drawn from the NAKO and COSYCONET cohorts to calculate 95% confidence intervals for the median liver and lung segmentation uncertainties of the PWD-based error estimators (Tab. C.2). Similarly, CIs were determined for the correlation coefficients of paired liver volume, fat, and lung volume measurements and their improvement through QC (Tab. C.3). Finally, CIs were determined for lung mask characteristics, including the P_{90} percentiles of small connected-component voxel counts and maximum hole sizes, as well as median sphericity in the retained and rejected groups (Tab. C.4). Cliff's δ was additionally determined between these groups, and group differences were assessed using two-sided Wilcoxon rank-sum tests with Holm-adjusted p values.

Table C.1: Quality control performance of all error estimators across participant-level bootstrap samples, with performance assessed by alignment with delineation accuracy, segmentation selection and failure case sampling capability, including corresponding 95% confidence intervals. E-AURC and $AUPRC_F$ values are reported as percentages for improved readability.

Dataset	Method	Pearson r	Binary E-AURC (%)	Continuous E-AURC (%)	$AUPRC_F$ (%)
2pt-Dixon (liver)	Region max-softmax	0.574 [0.195, 0.813]	3.77 [1.38, 6.57]	2.54 [1.10, 4.16]	58.4 [38.5, 75.7]
	Region entropy	0.572 [0.203, 0.812]	3.65 [1.30, 6.42]	2.52 [1.09, 4.14]	59.0 [38.9, 76.2]
	Coeff. of Var.	0.488 [0.274, 0.690]	0.85 [0.37, 1.61]	1.20 [0.95, 1.43]	56.5 [41.6, 68.8]
	VAE	0.667 [0.404, 0.842]	2.36 [0.76, 4.73]	1.37 [0.97, 1.75]	57.2 [35.4, 76.7]
	Quality Regressor	0.713 [0.536, 0.857]	2.53 [1.22, 4.34]	1.35 [1.04, 1.66]	48.2 [28.8, 65.2]
	Random Forest	0.644 [0.351, 0.794]	2.90 [1.06, 5.49]	1.36 [0.97, 1.81]	56.0 [36.8, 72.3]
	Mean PWD	0.707 [0.575, 0.853]	0.63 [0.26, 1.23]	1.12 [0.85, 1.34]	63.4 [50.0, 75.4]
	Median PWD	0.764 [0.631, 0.920]	0.44 [0.18, 0.84]	0.98 [0.74, 1.18]	75.4 [63.0, 84.7]
2pt-Dixon Test (liver)	Region max-softmax	0.308 [0.184, 0.428]	3.10 [0.52, 6.87]	0.72 [0.54, 0.91]	25.2 [16.7, 39.5]
	Region entropy	0.331 [0.193, 0.462]	2.85 [0.44, 6.39]	0.69 [0.51, 0.88]	25.6 [16.9, 40.3]
	Coeff. of Var.	0.338 [0.198, 0.465]	2.00 [0.56, 4.19]	0.64 [0.53, 0.73]	21.3 [15.0, 34.5]
	VAE	0.182 [−0.121, 0.438]	3.42 [1.28, 6.16]	0.74 [0.49, 1.01]	12.8 [6.0, 27.8]
	Quality Regressor	0.319 [0.059, 0.491]	3.30 [1.86, 5.61]	0.64 [0.45, 0.85]	8.9 [6.1, 16.0]
	Random Forest	0.394 [0.261, 0.550]	1.57 [0.19, 3.61]	0.73 [0.57, 0.87]	30.5 [22.0, 44.3]
	Mean PWD	0.548 [0.485, 0.605]	0.49 [0.14, 1.04]	0.39 [0.31, 0.45]	35.2 [28.4, 47.6]
	Median PWD	0.595 [0.503, 0.664]	0.49 [0.13, 1.08]	0.39 [0.31, 0.45]	39.8 [33.0, 50.1]
6pt-Dixon (liver)	Region max-softmax	0.605 [0.373, 0.763]	3.55 [1.18, 6.87]	2.28 [1.67, 2.94]	62.6 [35.9, 82.3]
	Region entropy	0.615 [0.389, 0.771]	3.44 [1.09, 6.78]	2.23 [1.64, 2.86]	63.5 [36.8, 83.1]
	Coeff. of Var.	0.664 [0.567, 0.794]	2.40 [0.95, 4.49]	2.20 [1.76, 2.60]	69.1 [47.2, 83.8]
	VAE	0.477 [0.172, 0.684]	5.47 [1.76, 10.65]	3.27 [2.45, 4.11]	55.5 [29.3, 76.2]
	Quality Regressor	0.375 [0.129, 0.583]	10.23 [3.88, 17.23]	3.68 [2.34, 5.32]	31.3 [18.3, 48.4]
	Random Forest	0.778 [0.602, 0.873]	5.82 [2.03, 11.10]	1.88 [1.36, 2.47]	57.9 [35.0, 77.1]
	Mean PWD	0.789 [0.709, 0.868]	1.52 [0.56, 2.94]	1.46 [1.19, 1.74]	75.5 [55.8, 88.5]
	Median PWD	0.762 [0.680, 0.858]	1.73 [0.63, 3.30]	1.34 [1.05, 1.66]	71.4 [51.0, 85.3]
2pt-Dixon (lung)	Region max-softmax	0.587 [0.308, 0.778]	4.76 [1.49, 9.00]	0.76 [0.48, 1.08]	44.4 [15.4, 68.5]
	Region entropy	0.584 [0.307, 0.770]	4.56 [1.35, 8.68]	0.73 [0.46, 1.04]	44.3 [14.4, 68.4]
	Coeff. of Var.	0.222 [0.068, 0.394]	1.85 [0.57, 3.77]	0.70 [0.54, 0.85]	28.2 [18.3, 42.2]
	VAE	0.743 [0.556, 0.869]	0.40 [0.11, 0.88]	0.78 [0.38, 1.02]	74.0 [64.4, 88.3]
	Quality Regressor	0.664 [0.320, 0.832]	2.70 [0.98, 5.12]	0.70 [0.41, 1.05]	53.7 [31.0, 77.7]
	Random Forest	0.752 [0.614, 0.851]	0.67 [0.26, 1.25]	0.33 [0.20, 0.44]	59.9 [39.6, 76.5]
	Mean PWD	0.509 [0.365, 0.623]	0.72 [0.26, 1.46]	0.58 [0.43, 0.70]	42.3 [27.8, 56.1]
	Median PWD	0.695 [0.540, 0.801]	0.40 [0.12, 0.94]	0.44 [0.33, 0.54]	63.6 [48.1, 79.0]
Fat-Sat (lung)	Region max-softmax	−0.399 [−0.666, −0.024]	32.80 [3.26, 62.98]	1.58 [0.72, 2.51]	20.9 [7.8, 42.2]
	Region entropy	−0.343 [−0.606, 0.031]	30.76 [2.93, 62.92]	1.54 [0.69, 2.52]	21.3 [8.3, 42.6]
	Coeff. of Var.	0.048 [−0.015, 0.189]	12.33 [1.00, 26.76]	0.69 [0.32, 1.07]	42.2 [20.0, 66.7]
	VAE	−0.312 [−0.708, 0.059]	41.10 [4.38, 65.12]	1.71 [0.71, 2.51]	16.6 [7.6, 33.3]
	Quality Regressor	0.316 [0.062, 0.707]	5.35 [1.31, 13.72]	0.56 [0.23, 0.79]	47.3 [18.8, 91.1]
	Random Forest	0.257 [−0.174, 0.670]	17.89 [1.14, 40.46]	0.78 [0.23, 1.37]	43.3 [12.8, 75.4]
	Mean PWD	0.377 [0.232, 0.694]	2.45 [0.36, 6.04]	0.20 [0.09, 0.41]	69.3 [43.2, 85.2]
	Median PWD	0.610 [0.443, 0.828]	1.99 [0.31, 5.06]	0.18 [0.08, 0.37]	77.0 [49.9, 91.0]

QC increased the correlations of paired downstream measurements in all three analyses, with 95% CIs of Δr excluding zero for NAKO liver volume, NAKO liver fat, and COSYCONET lung volume (Tab. C.3). The largest improvement was observed for NAKO liver fat, where the correlation increased from $r = 0.361$ to $r = 0.975$.

QC-retained and QC-rejected COSYCONET lung segmentations also differed significantly in mask characteristics (Tab. C.4). Rejected segmentations contained more small disconnected components and larger holes and showed lower sphericity than retained segmentations; all three differences remained significant after Holm correction.

Appendix C.3. Assessment of SOP conformance and segmentation performance evaluation

Second-rater manual liver segmentations and performance evaluations were performed in the 2pt-Dixon Test dataset, assessing the systematic influence of SOP-non-conformance in this dataset (Sects. 5.2 and 5.4 of the main article). Five random cases were sampled from the dataset and manual liver annotations were produced conforming with the described SOP. Subsequently, the cases were automatically segmented using the fully-trained liver segmentation ensemble for 2pt-Dixon data (Sect. 5.3 of the main article). The Jaccard index was determined for these segmentations, comparing scores against the original, non-SOP-conforming and the second-rater SOP-conforming liver annotations (Fig. C.2). On average, automated liver segmentations showed a 1.12% lower Jaccard similarity index (95% CI: 0.84, 1.50, based on $n = 10{,}000$ bootstrap resamples) compared with SOP-conforming annotations of the same cases. The observed performance bias is consistent with the selected Jaccard threshold indicating erroneous vs error-free segmentations in the 2pt-Dixon Test dataset (91% as compared to 92% for non-SOP- vs SOP-conforming 2pt-Dixon annotations, comp. Sect. 5.4 of the main article).

Appendix C.4. Sensitivity of segmentation selection to the Jaccard-based failure threshold

To assess the sensitivity of segmentation selection performance to the chosen Jaccard-based failure threshold, we repeated the AURC analysis across a range of threshold values (Fig. C.3). Across the evaluated Jaccard-based failure thresholds, the PWD-based AURC curves remained closest to the oracle curves, while the relative ranking of error estimators remained largely stable. This indicates that the observed advantage of PWD-based estimators was not an artifact of the selected heuristic failure threshold. At increasingly strict failure thresholds, attainable segmentation selection performance was naturally limited by the corresponding oracle curves, beyond which even an ideal error estimator could no longer achieve meaningful low-risk selection. Conversely, decreasing the failure threshold generally improved AURC values across all estimators.

Table C.2: Descriptive summary of PWD-based error estimates obtained during segmentation quality control in the NAKO and COSYCONET cohorts. Cohort-level medians are reported with 95% confidence intervals (CIs) based on 10,000 bootstrap resamples.

Cohort	Sequence	ROI	Estimator	Median uncertainty [95% CI]
NAKO	2pt-Dixon	Liver	Median PWD	0.00555 [0.00553, 0.00558]
NAKO	6pt-Dixon	Liver	Mean PWD	0.00846 [0.00840, 0.00851]
NAKO	2pt-Dixon	Lung	Median PWD	0.00347 [0.00346, 0.00348]
COSYCONET	FatSat	Lung	Median PWD	0.00571 [0.00538, 0.00600]

Table C.3: Alignment of paired measurements before and after QC selection. Pearson correlation coefficients (r) and their changes after QC selection (Δr) are reported with 95% confidence intervals (CIs) based on 10,000 bootstrap resamples.

Analysis	r before QC [95% CI]	r after QC [95% CI]	Δr [95% CI]
NAKO liver volume	0.968 [0.960, 0.975]	0.992 [0.991, 0.993]	0.024 [0.017, 0.033]
NAKO liver fat	0.361 [0.329, 0.396]	0.975 [0.958, 0.984]	0.614 [0.577, 0.647]
COSYCONET lung volume	0.866 [0.826, 0.897]	0.897 [0.874, 0.917]	0.031 [0.004, 0.067]

Table C.4: Lung mask characteristics of QC-retained and QC-rejected COSYCONET segmentations. Summary statistics and Cliff's δ effect sizes are reported with 95% confidence intervals (CIs) based on 10,000 bootstrap resamples. Statistical significance testing was performed using two-sided Wilcoxon rank-sum tests, with p values adjusted for multiple comparisons using the Holm method. Cliff's δ is defined as retained minus rejected.

Characteristic	Estimator	Retained estimate [95% CI]	Rejected estimate [95% CI]	Cliff's δ [95% CI]	Holm p
Small-component voxels	P_{90}	6.0 [4.0, 8.0]	135.8 [61.7, 272.9]	−0.524 [−0.629, −0.418]	3.04×10^{-19}
Largest hole size	P_{90}	2.0 [1.0, 2.5]	21.9 [2.0, 76.2]	−0.190 [−0.290, −0.089]	9.31×10^{-6}
Sphericity	Median	0.4969 [0.4952, 0.5000]	0.4750 [0.4709, 0.4791]	0.480 [0.365, 0.590]	9.50×10^{-14}

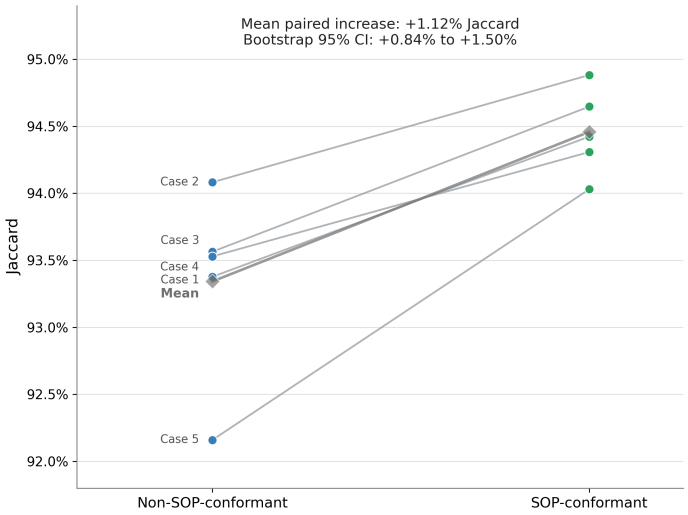


Figure C.2: **Non-conformance with liver segmentation SOP affects Jaccard index measurements:** Models trained on SOP-conforming manual annotations achieve systematically lower Jaccard index scores on non-SOP-conforming cases in the 2pt-Dixon Test dataset. On randomly selected cases within this dataset, they score 1.12% lower on average as compared to SOP-conforming annotations performed on the same cases. The observed performance bias is consistent with the selected Jaccard threshold indicating erroneous vs error-free segmentations in the 2pt-Dixon Test dataset (91% as compared to 92% for non-SOP- vs SOP-conforming 2pt-Dixon annotations).

Appendix C.5. Comparison with prior automated quality-control approaches

Tab. C.9 summarizes characteristics of representative automated approaches for quality control and error estimation in population-scale imaging. The comparison focuses on practical requirements for large-scale segmentation QC, including annotation demand, computational cost, ability to detect segmentation errors within the high-Dice range, support for segmentation selection and failure sampling, automated threshold optimization, and applicability to 3D segmentation. The proposed approach combines on-the-fly ensemble-based error estimation with segmentation selection, failure sampling, and automated threshold optimization while requiring no additional manual annotations for error estimation itself; the annotated cases indicated in parentheses were used for threshold optimization.

Appendix C.6. Visual analysis of silent failures

Fig. C.4 illustrates representative silent failures observed during error estimation. In these cases, substantial segmentation errors were assigned comparatively high confidence by individual error estimators and would therefore be included in selected segmentation sets at low coverage. The examples further show that ensemble output masks and voxel-wise entropy maps often captured uncertainty in the affected regions, but that the evaluated error estimators differed in how effectively this information was aggregated into a case-level uncertainty score. These examples support the interpretation of early rises in the corresponding RC curves as genuine silent failures rather than artifacts of the selected Jaccard-based failure threshold.

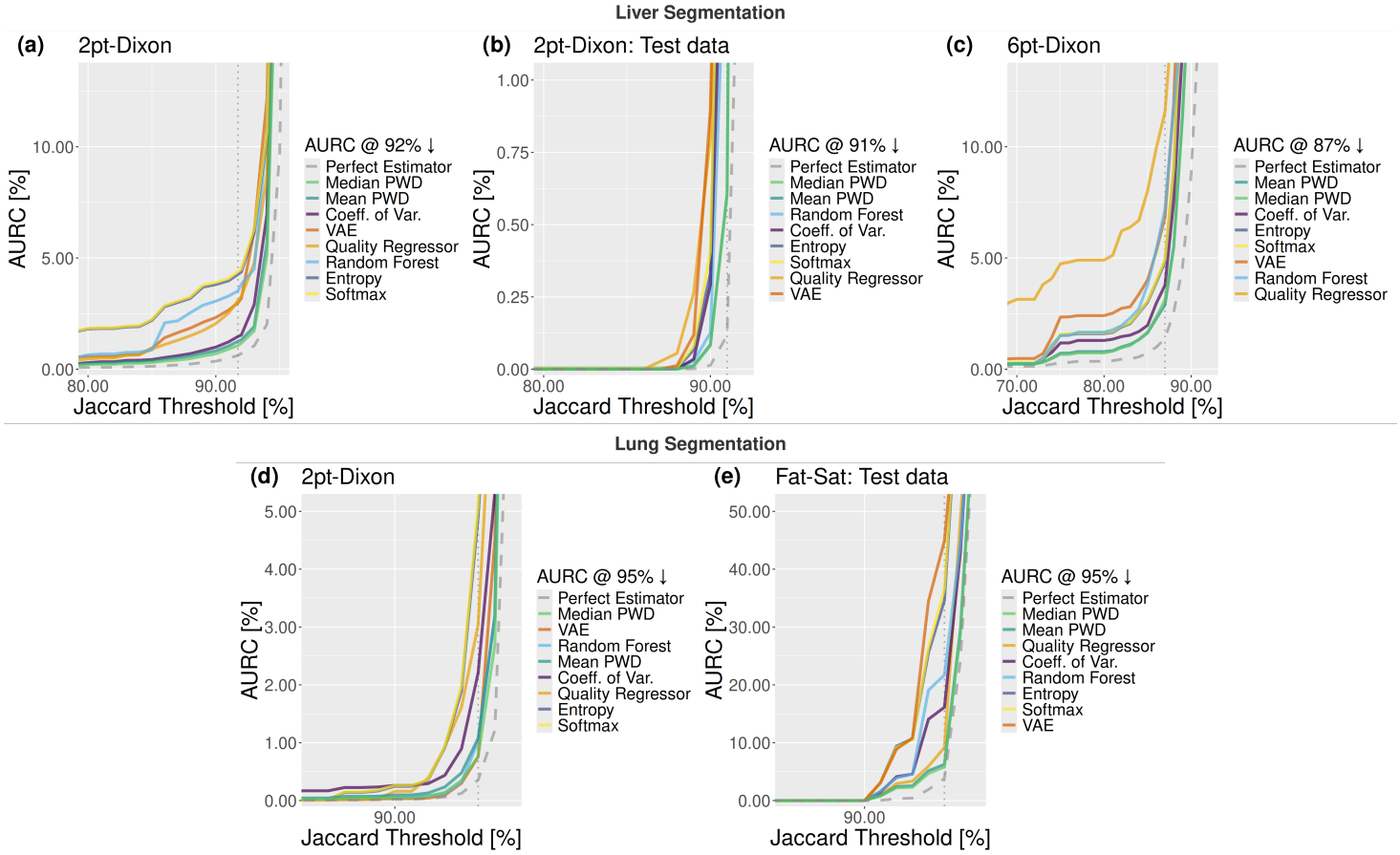


Figure C.3: **Segmentation selection performance across Jaccard-defined error levels:** Performance ranking based on the AURC (y-axis) of the evaluated error estimators remained largely stable across a wide range of Jaccard thresholds used to classify erroneous and error-free segmentations (x-axis). PWD-based estimators (green, cyan) performed well in selecting error-free segmentations across thresholds and most closely matched the performance of an ideal error estimator (dashed gray curve). The ideal estimator further indicated an upper bound on the Jaccard threshold, beyond which even a perfect error estimator fails to achieve meaningful segmentation selection (reflected by the near-vertical rise of gray curves). The manually selected Jaccard thresholds (dotted vertical lines) were generally close to these bounds and corresponded to non-optimistic AURC scores across all methods. With decreasing thresholds, all error estimators approached perfect AURC scores, indicating that overly lenient thresholds lose their ability to provide a meaningful measure of error for segmentation selection.

Table C.5: **Optimal uncertainty thresholds for selecting liver segmentations in 2pt-Dixon data:** thresholds were automatically determined to maximize the number of selected segmentations. At the same time, upper error risk bounds were ensured to meet predefined target risk levels. The resulting risk bounds closely aligned with the target levels. Observed error risks in both the training and additional test datasets were consistent, remaining well below the calculated bounds. All computations were performed with a confidence level exceeding 99% (Sect. 4.3 of the main article).

Target risk level [%]	Upper error risk bound [%]	Training risk [%]	Test risk [%]	Training coverage [%]	Test coverage [%]	NAKO coverage [%]	NAKO max expected errors	Selection threshold
1.00	0.92	0.00	0.22	71.46	57.46	98.37	101	0.0105
1.50	1.43	0.26	0.26	74.10	64.25	98.93	158	0.0112
2.00	1.82	0.49	0.77	76.84	75.33	99.36	202	0.0125
2.50	2.37	0.85	1.15	78.17	79.42	99.47	264	0.0132
3.00	2.95	1.28	1.84	81.47	88.25	99.63	329	0.0156
3.50	3.41	1.61	2.04	82.33	90.00	99.64	380	0.0164
4.00	3.98	2.04	2.49	83.55	92.08	99.71	444	0.0175
4.50	4.37	2.35	2.60	84.40	93.08	99.72	488	0.0181

Table C.6: **Optimal uncertainty thresholds for selecting liver segmentations in 6pt-Dixon data:** thresholds were automatically determined to maximize the number of selected segmentations. At the same time, upper error risk bounds were ensured to meet predefined target risk levels. The resulting risk bounds closely aligned with the target levels. In the 6pt-Dixon dataset, risk bounds lower than 2.50% could not be established. All computations were performed with a confidence level exceeding 99% (Sect. 4.3 of the main article).

Target risk level [%]	Upper error risk bound [%]	Training risk [%]	Training coverage [%]	NAKO coverage [%]	NAKO max expected errors	Selection threshold
2.50	2.44	0.00	47.35	97.32	266	0.0238
3.00	2.44	0.00	47.35	97.32	266	0.0238
3.50	3.17	0.35	48.72	97.35	345	0.0244
4.00	3.69	0.67	50.94	97.39	402	0.0252
4.50	4.15	0.97	52.65	97.42	452	0.0256
5.00	4.66	1.29	52.99	97.42	508	0.0258
5.50	5.10	1.59	53.85	97.43	556	0.0260
6.00	5.81	2.25	60.68	97.48	634	0.0284

Table C.7: **Optimal uncertainty thresholds for selecting lung segmentations in 2pt-Dixon data:** thresholds were automatically determined to maximize the number of selected segmentations. At the same time, upper error risk bounds were ensured to meet predefined target risk levels. The resulting risk bounds closely aligned with the target levels. Thresholding at the lowest target risk level of 1.6% ensured minimal error in the external test dataset. Beyond this level, the observed test risk increased beyond the expected error tolerance. All computations were performed with a confidence level exceeding 99% (Sect. 4.3 of the main article).

Target risk level [%]	Upper error risk bound [%]	Training risk [%]	Test risk [%]	Training coverage [%]	Test coverage [%]	NAKO coverage [%]	NAKO max expected errors	Selection threshold
1.60	1.59	0.00	0.00	56.24	25.41	99.48	177	0.0076
1.80	1.64	0.18	4.10	73.32	62.59	99.95	183	0.0106
2.00	1.88	0.34	10.46	77.66	75.76	99.96	210	0.0119
2.25	2.15	0.50	11.86	78.45	77.62	99.96	240	0.0122
2.50	2.39	0.66	14.02	80.03	81.47	99.97	268	0.0128
2.75	2.61	0.80	17.46	81.73	86.13	99.97	292	0.0139
3.00	2.83	0.95	18.82	82.79	88.58	99.98	317	0.0144
3.25	3.08	1.11	18.82	82.92	88.58	99.98	344	0.0144
3.50	3.47	1.40	20.69	84.76	91.26	99.99	388	0.0153

Table C.8: **Optimal uncertainty thresholds for selecting lung segmentations in the Cosyconet cohort:** thresholds were automatically determined to maximize the number of selected segmentations. At the same time, upper error risk bounds were ensured to meet predefined target risk levels. The resulting risk bounds closely aligned with the target levels. Risk bounds lower than 2.00% could not be established in this setting. All computations were performed with a confidence level exceeding 99% (Sect. 4.3 of the main article).

Target risk level [%]	Upper error risk bound [%]	Training risk [%]	Training coverage [%]	Cosyconet coverage [%]	Cosyconet max expected errors	Selection threshold
2.00	1.95	0.00	40.68	80.71	8	0.0087
2.50	2.34	0.26	44.87	83.15	10	0.0091
3.00	2.78	0.51	46.04	83.15	12	0.0091
3.50	3.25	0.76	46.15	83.15	14	0.0091
4.00	3.69	1.14	51.05	84.46	17	0.0095
4.50	4.35	1.67	55.83	85.39	20	0.0099
5.00	4.74	2.00	58.16	85.58	22	0.0101

Table C.9: **Comparison of automated quality-control (QC) techniques for computational results in population-based datasets:** The proposed method uniquely enables on-the-fly QC of 3D segmentations without requiring extensive data annotations, such as those typically used during error estimation training. This capability is not found in other approaches. To the best of our knowledge, it is also the only method demonstrated to be effective in selecting accurate segmentations, while providing an automated search for optimal selection thresholds. N/A = not applicable in non-segmentation contexts, or irrelevant to the objectives of the respective study.

Method	Manual Annotations for Err. Estim. (+ Thresh. Opt.)	Computation Time / Case [s]	Error Handling at > 0.90 Dice	Acc. Estim.	Segmentation Selection	Error Sampling	Thresh. Opt.	3D-Segm.
RCA[1]	100	660.00	No	Yes	No	No	No	Yes
CNN[2]	4,882	0.04	No	Yes	No	No	No	Yes
VAE[3]	1,000	'real-time'	No	Yes	No	No	No	Yes
Lin. Model[4]	2,121	0.39	No	Yes	No	No	No	No
Mean Var.[5]	8,539	0.02	N/A	N/A	N/A	N/A	No	No
PW-ASSD[6]	0	'on-the-fly'	No	Yes	No	No	N/A	No
Proposed	0 (+125)	0.92	Yes	Yes	Yes	Yes	Yes	Yes

[1] Robinson et al. (2019); [2] Robinson et al. (2018); [3] Wang et al. (2020); [4] Hann et al. (2021); [5] Langner et al. (2021); [6] Ng et al. (2022).

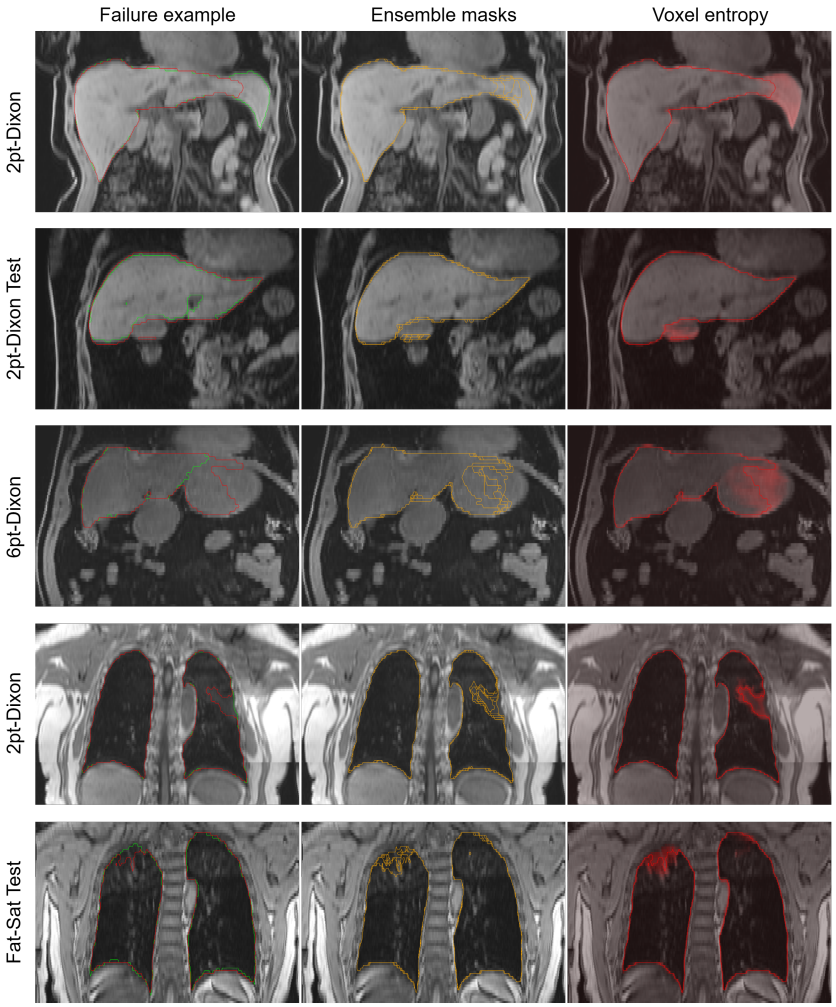


Figure C.4: **Examples of silent failures during error estimation:** Various segmentation failures (column 1; red: automatic segmentations, green: ground truth) were assigned high confidence by non-PWD error estimators across datasets (top to bottom). Ensemble output masks (column 2) and voxel-wise entropy maps (column 3) capture uncertainty over the same error regions using distinct representations. The evaluated estimators differed in their ability to effectively leverage this information during QC.

References

Geifman, Y., El-Yaniv, R., 2017. Selective classification for deep neural networks. Advances in neural information processing systems 30.

Hann, E., Popescu, I.A., Zhang, Q., Gonzales, R.A., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K., 2021. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. Medical image analysis 71, 102029.

Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation 14, 282.

Langner, T., Gustafsson, F.K., Avelin, B., Strand, R., Ahlström, H., Kullberg, J., 2021. Uncertainty-aware body composition analysis with deep regression ensembles on UK Biobank MRI. Computerized Medical Imaging and Graphics 93, 101994.

Ng, M., Guo, F., Biswas, L., Petersen, S.E., Piechnik, S.K., Neubauer, S., Wright, G., 2022. Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study. IEEE Transactions on Biomedical Engineering .

Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Kainz, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Rueckert, D., Glocker, B., 2018. Real-time prediction of segmentation quality, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. volume 11073, pp. 578–585.

Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Matthews, P.M., Rueckert, D., Glocker, B., 2019. Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance 21, 18.

Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., Bai, W., 2020. Deep generative model-based quality control for cardiac MRI segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. volume 12264, pp. 88–97.