Beyond Dice: Risk-Normalized and Hazard-Aware Evaluation of Medical Segmentation for Image-Guided Robotics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Safety in embodied systems depends on where perception fails, not only on how often it fails. We introduce two bounded evaluation metrics for segmentation that make spatial risk explicit using an anatomy-derived hazard field built from distance to protected structures. The *Safety Impact Score* (SIS) measures the share of total hazard mass that is misclassified, with a tunable trade-off between false negatives and false positives. The *Safety Tail Risk* (STAR) summarizes the worst fraction of error hazards using a conditional value-at-risk operator. To isolate metric behavior from model quality, we design a model-free matched-Dice stress test that relocates equal numbers of boundary errors toward or away from hazard while keeping Dice unchanged. We run this protocol on three public Medical Segmentation Decathlon tasks (Hepatic Vessel, Liver, Pancreas; five cases each). Across datasets, STAR shows large positive deltas for the risky variant (combined mean $\Delta\text{STAR} = 0.431 \pm 0.124$, paired Wilcoxon $p = 3.24\times10^{-4}$), and SIS is also positive (combined mean $\Delta\text{SIS} = 0.0818 \pm 0.0771$, $p = 3.05\times10^{-5}$). Effects are strongest when the hazard corresponds to vessels in Hepatic Vessel (mean $\Delta\text{STAR} = 0.614 \pm 0.307$, mean $\Delta\text{SIS} = 0.224 \pm 0.182$). A proximity-weighted overlap baseline (hazard-weighted Dice) moves little or in the opposite direction. Results persist with an exponential hazard kernel, indicating robustness. These findings demonstrate that risk-normalized and tail-aware evaluation captures safety-relevant differences that overlap metrics miss, using only public data and a simple perturbation protocol.

1 Background

1.1 Why average segmentation accuracy is not safety

Perception modules increasingly guide actions in image–guided and robotic procedures, where the severity of an error depends on *where* it occurs. A small miss that pushes a boundary toward a critical structure can be far more consequential than a larger miss in benign tissue. Conventional metrics such as Dice, Jaccard, and boundary distances treat all voxels or boundary points roughly uniformly and typically report means or percentiles aggregated over space and cases. Multiple studies show that these scores can fail to predict downstream harm or workload. In radiotherapy, geometric scores correlate weakly with dose deviations and clinical acceptability, which motivates multi–aspect evaluation that includes treatment impact and expert ratings rather than geometry alone Poel et al. (2021, 2025); Maier-Hein et al. (2024); Reinke et al. (2023). Work on edit–effort and clinical usability points the same way: Added Path Length, surface–tolerant measures such as Normalized Surface Dice, and the Mendability Index track how hard clinicians must work to fix outputs, not just how

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

similar masks look He et al. (2024); Various (2024); Ma et al. (2024); He et al. (2023); Zhang et al. (2024). Average overlap is therefore a poor proxy for *risk*.

1.2 What has already been done to inject safety into metrics

Proximity– or sensitivity–weighted overlap in radiotherapy. Several groups weight overlap by anatomical importance. OAR–DSC reweights Dice by proximity to organs at risk and by radiosensitivity McCullum et al. (2024). Weighted Dice for brachytherapy similarly uses distance to high–risk volumes and reports stronger links to dose endpoints Ni et al. (2025). These are important steps toward task–aware evaluation.

Outcome–aware validation frameworks. Comprehensive frameworks evaluate contouring by geometry, efficiency, plan quality, and expert acceptability, which exposes outliers missed by average metrics Poel et al. (2025). Guidance now recommends reporting complementary measures and considering task risk explicitly Maier-Hein et al. (2024); Reinke et al. (2023).

Tail–sensitive boundary metrics. Hausdorff distance and HD95 emphasize extreme boundary errors Huttenlocher et al. (1993); Karimi & Salcudean (2019). They are useful for spotting worst–case deviations, yet agnostic to which anatomy those deviations threaten.

Safety–aware perception outside medicine. In autonomous driving, safety metrics for semantic segmentation penalize clustered errors and errors in safety–critical regions instead of averaging over all pixels Cheng et al. (2021). Task–aware risk estimators quantify how perception failures propagate into risky plans and highlight low–probability high–impact events Antonante et al. (2023).

Risk modeling around vital anatomy for robotic procedures. Surgical robotics often frames safety as distance–to–hazard with safety margins that expand where uncertainty is higher. Inner ear and bone–milling work formalize probabilistic error budgets and turn them into spatial keep–out regions around critical structures Dillon et al. (2016); Siebold et al. (2017). This connects naturally to evaluating perception in terms of *clearance* to protected anatomy.

1.3 Where the gaps remain

- **Granular spatial risk in the metric itself.** Existing weighted Dice variants attach scalar weights to overlap, but they do not evaluate errors through a continuous hazard field that reflects millimeter–scale clearance to protected structures. HD95 is tail–sensitive but not hazard–aware.
- **Risk–mass normalization for cross–case comparability.** Most proposals reweight a score rather than compute the *share of available risk* that the model gets wrong. Without normalizing by total hazard mass, the same numeric change can mean very different things across patients or scanners.
- **False negative versus false positive asymmetry.** Few evaluation metrics expose explicit cost trade–offs between risk–weighted FN and FP, even though misses near critical anatomy are often costlier than conservative over–segmentations.
- **Tail risk over hazard, not just over distance.** There is no standard evaluation that aggregates the *worst fraction of errors by hazard* in the spirit of conditional value at risk. Percentile Hausdorff is a boundary percentile, not a hazard–weighted tail of clinically dangerous errors.

1.4 How our approach is different

We introduce a small, principled family of evaluation metrics that make safety the first–class quantity.

Spatial hazard map from anatomy. We build a continuous hazard field $w(x) \in [0, 1]$ from distance to critical structures, with polynomial or exponential decay. The field can combine multiple organs at risk via conservative max aggregation or additive accumulation.

Risk–normalized error shares. Let $y(x) \in \{0,1\}$ be ground truth for the target, $\hat{y}(x) \in \{0,1\}$ a prediction, and $w(x)$ the hazard map. Define

$$\text{R-FN} \;=\; \frac{\sum_x \mathbf{1}\{y(x)=1,\ \hat{y}(x)=0\}\, w(x)}{\sum_x \mathbf{1}\{y(x)=1\}\, w(x)}, \quad \text{R-FP} \;=\; \frac{\sum_x \mathbf{1}\{y(x)=0,\ \hat{y}(x)=1\}\, w(x)}{\sum_x \mathbf{1}\{y(x)=0\}\, w(x)},$$

and the Safety Impact Score

$$\text{SIS}_\lambda \;=\; \lambda\,\text{R-FN} + (1-\lambda)\,\text{R-FP}, \qquad \lambda \in [0,1].$$

SIS is bounded in $[0,1]$, interpretable across cases, and exposes FN versus FP asymmetry through λ. This differs from proximity–weighted Dice, which remains an overlap score and does not separate FN from FP or normalize by total risk McCullum et al. (2024); Ni et al. (2025).

Hazard–aware tail emphasis. Let $E_{\text{FN}} = \{x :\ y=1,\ \hat{y}=0\}$, $E_{\text{FP}} = \{x :\ y=0,\ \hat{y}=1\}$, and $h(x) = w(x)$. For $\alpha \in (0,1]$, let $q_{1-\alpha}$ be the $(1-\alpha)$ quantile of $h(x)$ over a set E and define $\text{CVaR}_\alpha(h \mid E) = \frac{1}{\alpha|E|}\sum_{x\in E} h(x)\,\mathbf{1}\{h(x) \geq q_{1-\alpha}\}$. The Safety Tail Risk is

$$\text{STAR}_{\lambda,\alpha} \;=\; \lambda\,\text{CVaR}_\alpha(h \mid E_{\text{FN}}) \;+\; (1-\lambda)\,\text{CVaR}_\alpha(h \mid E_{\text{FP}}).$$

STAR targets rare but catastrophic mistakes near protected anatomy, something boundary percentiles and region heuristics do not directly capture Cheng et al. (2021); Huttenlocher et al. (1993).

Theory and invariants. Because SIS and STAR integrate over a bounded hazard measure in millimeters, they are physically meaningful, invariant to voxel anisotropy when distances are in millimeters, and monotone when errors are moved closer to hazard. With a uniform hazard $w(x) \equiv 1$, SIS reduces to a standard weighted error rate, which shows that our construction strictly generalizes common accuracy measures.

Reproducible evidence. We provide a matched–Dice stress protocol and open–dataset scripts that move equal numbers of boundary voxels toward or away from hazard. Across three MSD tasks, SIS and especially STAR separate risky from neutral variants while Dice remains unchanged and hazard–weighted Dice moves little. This complements outcome–focused frameworks by offering a fast, anatomy–aware safety readout that can be run alongside classical metrics Poel et al. (2025).

1.5 Relationship to prior art

Our use of a distance–derived hazard field connects to safety margins in surgical robotics Dillon et al. (2016); Siebold et al. (2017). Our risk normalization and FN versus FP trade–off address long–standing complaints about average scores and symmetric penalties in clinical validation Maier-Hein et al. (2024); Poel et al. (2021). Our tail operator plays a role similar to worst–case planners in autonomous systems but applied to perception evaluation rather than control Antonante et al. (2023). Finally, proximity–weighted Dice becomes a special case of our framework under a binary ring–shaped hazard and symmetric weighting, which clarifies how our metrics generalize and strictly extend that line of work McCullum et al. (2024); Ni et al. (2025).

Summary. Prior work has weighted overlap by importance, emphasized boundary tails, and evaluated clinical impact, but there is still no bounded, hazard–aware metric that (i) computes risk–normalized FN and FP shares with explicit asymmetry and (ii) aggregates the worst errors by *hazard* rather than by distance or region size. SIS and STAR address these gaps with a continuous anatomy–derived hazard map, clear interpretation, simple properties, and reproducible demonstrations on public datasets.

2 Methodology

2.1 Problem setting and notation

Let $\Omega \subset \mathbb{Z}^3$ be a voxel grid with physical spacing (s_x, s_y, s_z) in millimeters and voxel volume $v = s_x s_y s_z$. Let $y : \Omega \to \{0,1\}$ be the ground truth mask of a target structure and $\hat{y} : \Omega \to \{0,1\}$ a prediction. Let $\{C_i\}_{i=1}^M$ denote critical structures to be protected, each provided as a binary mask. Errors are not equal: a false negative near C_i is typically more consequential than one far away. Our goal is to evaluate $\hat{y}$ with respect to y through a spatially varying hazard field derived from clearance to $\{C_i\}$.

2.2 Anatomy-derived hazard field

For each critical structure C_i, we compute the Euclidean distance transform in millimeters, $d_{C_i}(x)$, using an anisotropy-aware transform with sampling (s_x, s_y, s_z). We map distance to a unit hazard via a monotone decay $r : \mathbb{R}_{\geq 0} \to [0, 1]$:

$$\text{Polynomial clip:} \quad r(d) = \max\bigl\{0,\, 1 - (d/m)^p\bigr\}, \quad m > 0,\; p \geq 1, \tag{1}$$

$$\text{Exponential:} \quad r(d) = \exp(-d/\tau), \quad \tau > 0. \tag{2}$$

We combine multiple structures with either a conservative maximum or an additive accumulation clipped to one:

$$w(x) = \begin{cases} \max_i \; w_i \, r(d_{C_i}(x)) & \text{max aggregation,} \\ \min\Bigl\{1,\, \sum_i w_i \, r(d_{C_i}(x))\Bigr\} & \text{sum aggregation,} \end{cases} \tag{3}$$

where $w_i > 0$ are optional per-structure importances. The hazard map $w : \Omega \to [0, 1]$ reflects clearance to protected anatomy and serves as a spatial weight for evaluation.

2.3 Risk-normalized error shares and Safety Impact Score

Define risk-weighted false negative and false positive shares by normalizing the hazard mass of each error set by the total available hazard mass in the corresponding region:

$$\text{R-FN} = \frac{\sum_{x \in \Omega} \mathbf{1}\{y(x) = 1,\, \hat{y}(x) = 0\}\, w(x)}{\sum_{x \in \Omega} \mathbf{1}\{y(x) = 1\}\, w(x) + \varepsilon}, \tag{4}$$

$$\text{R-FP} = \frac{\sum_{x \in \Omega} \mathbf{1}\{y(x) = 0,\, \hat{y}(x) = 1\}\, w(x)}{\sum_{x \in \Omega} \mathbf{1}\{y(x) = 0\}\, w(x) + \varepsilon}, \tag{5}$$

with a small ε for numerical stability. We mix them with a tunable asymmetry $\lambda \in [0, 1]$:

$$\text{SIS}_\lambda \;=\; \lambda\, \text{R-FN} + (1 - \lambda)\, \text{R-FP}. \tag{6}$$

$\text{SIS}_\lambda \in [0, 1]$ is interpretable as the fraction of hazard mass that is misclassified, with explicit control of the false negative versus false positive trade off.

2.4 Hazard-aware tail emphasis and Safety Tail Risk

Let $E_{\text{FN}} = \{x \in \Omega :\, y = 1,\, \hat{y} = 0\}$ and $E_{\text{FP}} = \{x \in \Omega :\, y = 0,\, \hat{y} = 1\}$. Consider the per-error hazard values $h(x) = w(x) \in [0, 1]$. For $\alpha \in (0, 1]$, define the conditional value at risk over a set E:

$$\text{CVaR}_\alpha(h \mid E) = \frac{1}{\alpha\, |E|} \sum_{x \in E} h(x)\, \mathbf{1}\{h(x) \geq q_{1-\alpha}(E)\}\,, \tag{7}$$

where $q_{1-\alpha}(E)$ is the $(1 - \alpha)$ quantile of $\{h(x) : x \in E\}$. The Safety Tail Risk aggregates the worst fraction of error hazards for both error types:

$$\text{STAR}_{\lambda,\alpha} = \lambda\, \text{CVaR}_\alpha(h \mid E_{\text{FN}}) + (1 - \lambda)\, \text{CVaR}_\alpha(h \mid E_{\text{FP}})\,. \tag{8}$$

$\text{STAR}_{\lambda,\alpha} \in [0, 1]$ focuses on rare but catastrophic mistakes near protected anatomy.

2.5 Properties

Boundedness. Since $h \in [0, 1]$ and normalizers are strictly positive, $\text{R-FN}, \text{R-FP} \in [0, 1]$ and $\text{SIS}_\lambda \in [0, 1]$. Similarly, $\text{STAR}_{\lambda,\alpha} \in [0, 1]$.

Monotonicity. Moving any error voxel to a location with larger hazard $w(x)$ weakly increases SIS_λ and $\text{STAR}_{\lambda,\alpha}$, with denominators fixed.

Reduction. With $w \equiv 1$, SIS_λ reduces to a standard weighted error rate that generalizes voxel accuracy. STAR reduces to a tail average over uniform hazard.

Physical invariance. When distances are computed in millimeters, results are invariant to voxel anisotropy.

2.6 Baselines and comparators

We report standard and safety-aware baselines:

- Dice coefficient and HD95 as classical overlap and tail-boundary baselines.
- Hazard-weighted Dice (wDice): define

$$\mathrm{wDice} = \frac{2 \sum_x w(x)\, y(x)\, \hat{y}(x)}{\sum_x w(x)\, y(x) + \sum_x w(x)\, \hat{y}(x) + \varepsilon},$$

  which reweights overlap by w for comparison with overlap-style metrics.
- Proximity-weighted Dice variants in radiotherapy when applicable, treated as external baselines.

2.7 Matched-Dice stress protocol

To demonstrate added value beyond overlap, we design a model-free stress test that holds Dice approximately constant while relocating the same number of boundary voxels toward versus away from hazard.

1. Given a ground truth mask y and hazard map w, compute the inner boundary voxels B_{in} and outer boundary voxels B_{out} using a one-voxel erosion or morphological gradient.
2. Rank $B_{\mathrm{in}} \cap \{y = 1\}$ and $B_{\mathrm{out}} \cap \{y = 0\}$ by $w(x)$.
3. For a chosen k, construct two predictions:
   - risk-seeking: flip k highest-hazard inner boundary voxels to false negatives and k highest-hazard outer boundary voxels to false positives,
   - risk-neutral: flip k lowest-hazard inner boundary voxels to false negatives and k lowest-hazard outer boundary voxels to false positives.
4. Report Dice, HD95, wDice, SIS_λ, and $\mathrm{STAR}_{\lambda,\alpha}$ for both variants.

This protocol forces classical scores to move minimally while safety-aware metrics diverge when errors hug the hazard.

2.8 Datasets and label mappings

We use three public MSD tasks for generalization:

- **Task08 Hepatic Vessel** target is tumor (label 2) and hazard is vessels (label 1).
- **Task03 Liver** target is tumor (label 2) and hazard is liver (label 1).
- **Task07 Pancreas** target is tumor (label 2) and hazard is pancreas (label 1).

For each case, we read spacing from NIfTI headers to compute millimeter distances. We construct w with a polynomial kernel by default with $m = 10$ mm and $p = 2$, and also report an exponential kernel sensitivity with $\tau = 8$ mm.

2.9 Implementation details

Distances and morphology. We compute $d_{C_i}(x)$ via an anisotropy-aware Euclidean distance transform with sampling (s_x, s_y, s_z). Boundary sets use a 3D ball structuring element of radius 1 voxel. We ensure the same number of flipped voxels for risky and neutral variants at each case.

Numerics. All sums are carried out in 32-bit float. We use $\varepsilon = 10^{-8}$ in denominators. Quantiles for CVaR are computed with linear interpolation. We verify invariance to regridding by resampling select cases to isotropic spacing and confirming metrics change within tolerance.

Hyperparameters. Unless stated otherwise, we set $\lambda = 0.7$ to overweight false negatives and $\alpha = 0.05$ to summarize the top 5 percent most hazardous mistakes. We ablate $m \in \{5, 10, 15, 20\}$ mm, $p \in \{1, 2, 4\}$, $\tau \in \{6, 8, 10\}$ mm, and $\alpha \in \{0.01, 0.05, 0.1\}$.

Dataset	n	mean ΔSIS $\pm$ CI	p (SIS)	mean ΔSTAR $\pm$ CI	p (STAR)	mean ΔwDice
Liver	5	0.008742 ± 0.013258	0.03125	0.368505 ± 0.129822	0.03125	-0.087578
Pancreas	5	0.012759 ± 0.009152	0.03125	0.310152 ± 0.010858	0.03125	-0.117252
H. Vessel	5	0.223996 ± 0.181814	0.03125	0.614249 ± 0.307434	0.03125	-0.421403
Combined	15	0.0818 ± 0.0771	3.052×10^{-5}	0.4310 ± 0.1240	3.242×10^{-4}	–

Table 1: Effect sizes and significance for the matched–Dice stress test (polynomial kernel). CI denotes 95 percent confidence interval on the mean. p values from paired Wilcoxon signed–rank tests with alternative risky $>$ neutral. wDice reported as mean Δ across cases.

Multiple hazards. For multiple OARs we use max aggregation by default as a conservative choice and report sum aggregation as a sensitivity analysis.

2.10 Statistical analysis

We perform paired tests between risky and neutral variants within each dataset. For each metric we compute per-case deltas and report mean with 95 percent confidence intervals. We use a paired Wilcoxon signed-rank test with one-sided alternative that risky exceeds neutral for SIS and STAR. We aggregate across datasets and also report per-dataset results. For model-based settings, we compare methods across seeds and report bootstrap confidence intervals where appropriate.

3 Results

3.1 Experimental setup recap

We evaluate the proposed metrics using a matched–Dice stress protocol that holds overlap approximately constant while relocating equal numbers of boundary voxels toward versus away from hazard. We run on three Medical Segmentation Decathlon tasks with five cases each: Hepatic Vessel (target tumor, hazard vessels), Liver (target tumor, hazard liver), and Pancreas (target tumor, hazard pancreas). Unless stated otherwise we use a polynomial hazard kernel with margin m=10 mm and exponent p=2, tail fraction α=0.05, and asymmetry λ=0.7.

3.2 Primary finding: safety metrics separate risky from neutral while Dice does not

Across all datasets the tail metric $\Delta\mathrm{STAR} \equiv \mathrm{STAR}_{\mathrm{risky}} - \mathrm{STAR}_{\mathrm{neutral}}$ is strongly positive and statistically significant, indicating that equal–count errors placed nearer to hazard are penalized far more than those pushed away. The risk–normalized ΔSIS is also positive for all datasets, with the largest separation on Hepatic Vessel. By design $\Delta\mathrm{Dice}$ is zero and provides no safety discrimination, so we do not plot it.

3.3 Overlap baselines move little even when risk changes

We compare against a hazard–weighted Dice baseline (wDice), which reweights overlap by the hazard field. Although wDice shows some signal, it is substantially less consistent and smaller in magnitude than STAR.

3.4 Quantitative summary

Table 1 reports effect sizes and paired tests for the polynomial kernel. Hepatic Vessel shows the largest separation, consistent with the clinical intuition that tumors approaching vessels carry high risk. Combined across datasets, both ΔSIS and ΔSTAR are significant.

For completeness, Table 2 lists per–dataset means of risky and neutral metrics. STAR shows large absolute differences even when Dice is unchanged.

3.5 Sensitivity to hazard kernel choice

We repeat the analysis with an exponential kernel (τ=8 mm). Direction and relative magnitudes are preserved, indicating robustness to the kernel parameterization.

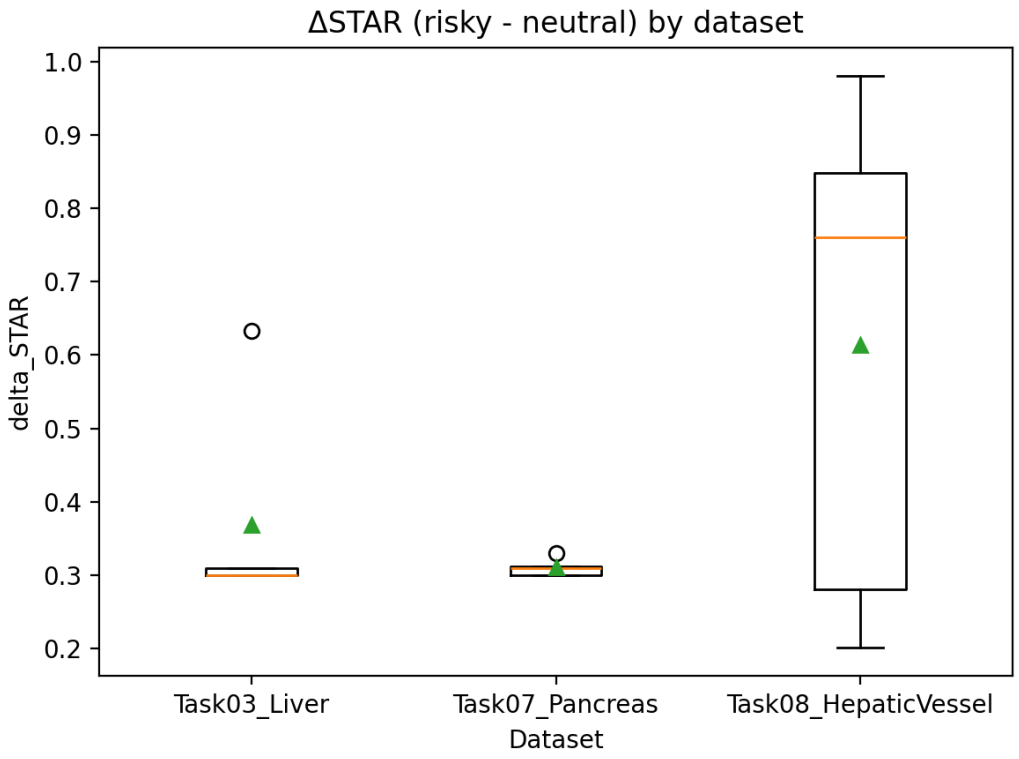

Figure 1: Per–dataset boxplots of ΔSTAR (risky minus neutral). Larger is worse for safety. All three datasets show a consistent positive gap, with the largest effect on Hepatic Vessel where vessels are an explicit organ at risk.

Dataset	Dice_r	Dice_n	SIS_r	SIS_n	STAR_r	STAR_n
Task03 Liver	0.6884	0.6884	0.2293	0.2205	0.9965	0.6280
Task07 Pancreas	0.7744	0.7744	0.1970	0.1843	0.9953	0.6852
Task08 HepaticVessel	0.6556	0.6556	0.4205	0.1965	0.8443	0.2300

Table 2: Per–dataset means for risky (r) and neutral (n) variants under the polynomial kernel. Dice is unchanged by construction while SIS and STAR separate.

3.6 Qualitative exemplars

To illustrate what the hazard field encodes, Figure 4 shows the hazard maps for Hepatic Vessel and Pancreas. These fields convert millimeter clearance to a spatial risk prior that aligns with clinical intuition and explains the observed metric behavior. We omit Task03 Liver exemplars since they are uninformative in this setup.

3.7 Observed edge cases and interpretation

On several Hepatic Vessel cases the neutral variant attains STAR near zero under the clipped polynomial kernel because neutral flips are pushed beyond the 10 mm margin where hazard weights clip to zero. This behavior is expected with a hard margin. The exponential kernel removes clipping and yields nonzero neutral STAR while preserving direction and magnitude, as in Table 3.

Summary SIS and STAR consistently penalize error relocations toward hazard across three datasets while Dice remains flat and wDice moves little. The effect is largest when the hazard corresponds to true organs at risk, and it persists under alternate hazard kernels. These results support the thesis that risk–normalized and tail–focused evaluation captures safety–relevant differences that overlap metrics miss.

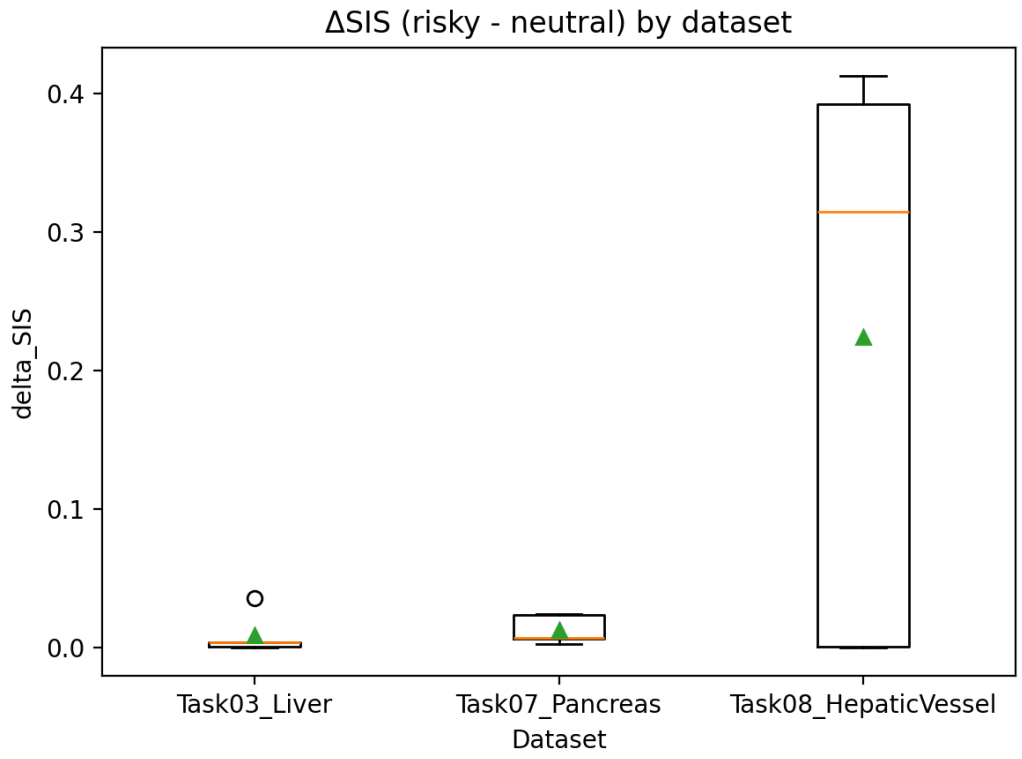

Figure 2: Per–dataset boxplots of ΔSIS (risky minus neutral). SIS separates risky from neutral on all datasets, again with the largest gap on Hepatic Vessel.

Dataset	mean ΔSIS (exp)	mean ΔSTAR (exp)
Task03 Liver	0.017233	0.379515
Task07 Pancreas	0.016671	0.341277
Task08 HepaticVessel	0.092108	0.472684

Table 3: Kernel sensitivity with exponential hazard. Effects remain positive and of similar order.

References

Pasquale Antonante, Sushant Veer, Karen Leung, Xinshuo Weng, Luca Carlone, and Marco Pavone. Task-aware risk estimation of perception failures for autonomous vehicles. In *Robotics: Science and Systems (RSS)*, 2023. URL https://www.roboticsproceedings.org/rss19/p100.pdf.

Chih-Hong Cheng, Alois Knoll, and Hsuan-Cheng Liao. Safety metrics for semantic segmentation in autonomous driving. *arXiv preprint arXiv:2105.10142*, 2021. URL https://arxiv.org/abs/2105.10142.

Neal P. Dillon, Michael A. Siebold, et al. Increasing safety of a robotic system for inner ear surgery using probabilistic error modeling near vital anatomy. In *Proc. SPIE Medical Imaging*, 2016. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5708556/.

Da He, Jayaram K. Udupa, Yubing Tong, and Drew A. Torigian. Mendability index: A new metric for estimating the effort required for manually editing auto-segmentations of objects of interest. In *Proc. SPIE Medical Imaging*, volume 12469, pp. 1246905, 2023. doi: 10.1117/12.2654421. URL https://pubmed.ncbi.nlm.nih.gov/37256076/.

Da He et al. Predicting the effort required to manually mend auto-segmentations in radiotherapy. *medRxiv*, 2024. URL https://www.medrxiv.org/content/10.1101/2024.06.12.24308779v1. Analysis of APL, surface Dice, and Mendability Index vs. human effort.

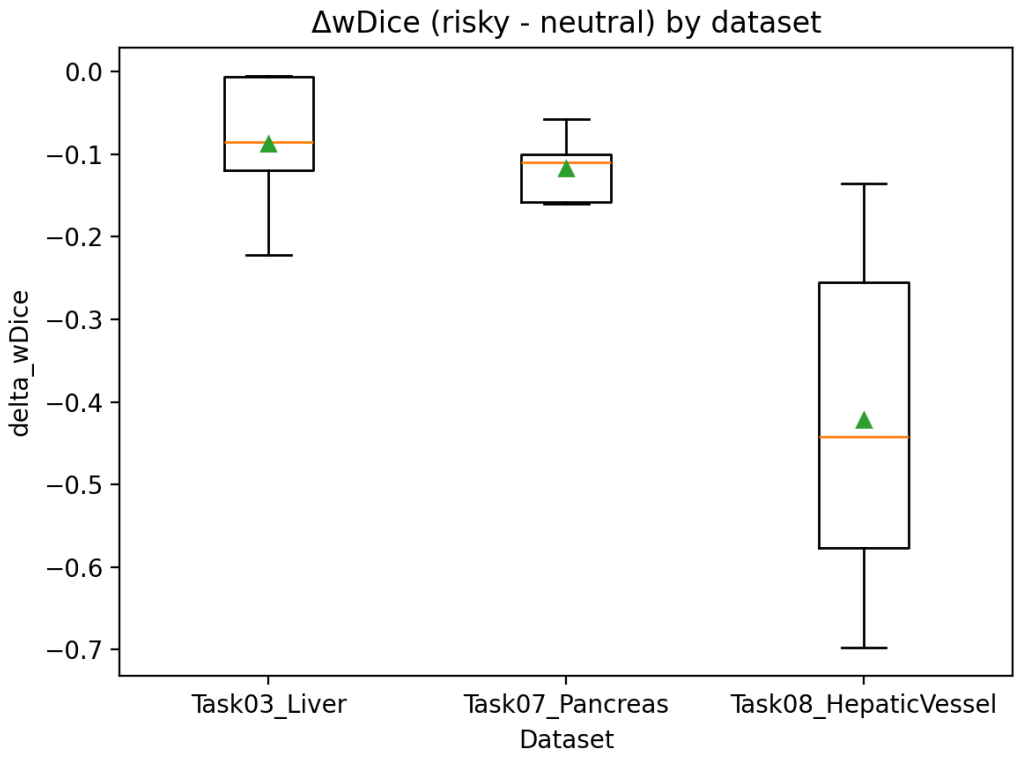

Figure 3: Per–dataset boxplots of ΔwDice (risky minus neutral). Values are close to zero with high variability compared to ΔSTAR. Negative means the risky variant scores worse under wDice.

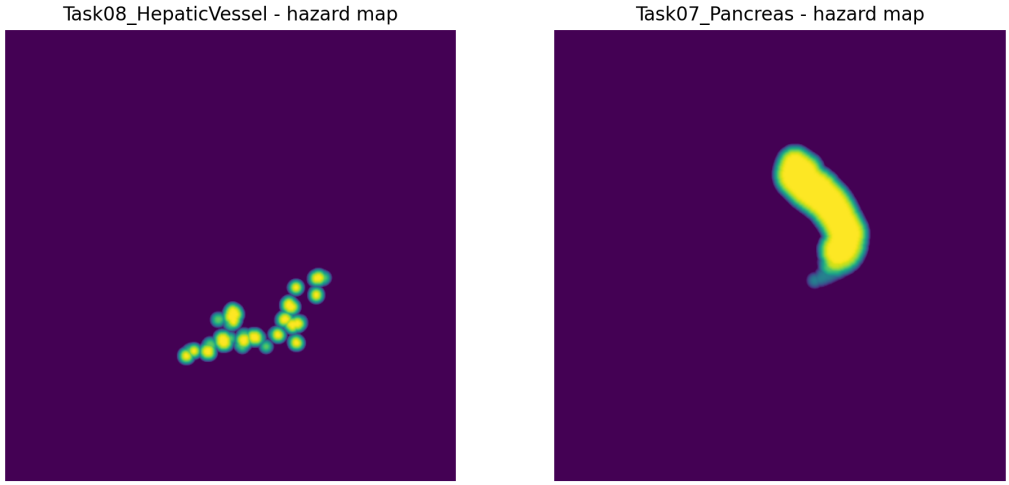

Figure 4: Hazard maps derived from distance to protected anatomy. Left: Hepatic Vessel (vessels as hazard). Right: Pancreas (organ as hazard). Brighter indicates higher hazard near critical structures.

Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. doi: 10.1109/34.232073. URL `https://dl.acm.org/doi/10.1109/34.232073`.

Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:1904.10030*, 2019. URL `https://arxiv.org/abs/1904.10030`.

Jun Ma et al. Segment anything in medical images. *Nature Communications*, 15(1160), 2024. URL https://www.nature.com/articles/s41467-024-44824-z. Uses NSD alongside DSC per Metrics Reloaded guidance.

Lena Maier-Hein, Annika Reinke, Patrick Godau, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z. URL https://www.nature.com/articles/s41592-023-02151-z.

Lucas McCullum, Kareem A. Wahid, Barbara Marquez, and Clifton D. Fuller. Oar-weighted dice score: A spatially aware, radiosensitivity aware metric for target structure contour quality assessment. *arXiv preprint arXiv:2410.20243*, 2024. URL https://arxiv.org/abs/2410.20243.

Rui Ni et al. Geometrically focused training and evaluation of organs-at-risk segmentation via deep learning. *Medical Physics*, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12257911/. Proposes weighted Dice (wDSC) using distance to high-risk CTV; reports stronger correlation with dosimetry.

Rianne Poel et al. The predictive value of segmentation metrics on dosimetry in organs at risk of the brain. *Medical Image Analysis*, 73:102161, 2021. doi: 10.1016/j.media.2021.102161. URL https://pubmed.ncbi.nlm.nih.gov/34293536/.

Rianne Poel et al. A comprehensive multifaceted technical evaluation framework for auto-segmentation in radiotherapy. *Communications Medicine*, 2025. URL https://www.nature.com/articles/s43856-025-01048-6. Evaluation across geometry, efficiency, treatment quality, and expert acceptability.

Annika Reinke et al. Understanding metric-related pitfalls in image analysis validation. *arXiv preprint arXiv:2302.01790*, 2023. URL https://arxiv.org/abs/2302.01790.

Michael A. Siebold, Loris Fichera, Neal P. Dillon, and J. Michael Fitzpatrick. Safety margins in robotic bone milling: from registration uncertainty to statistically safe surgeries. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 13(4):e1773, 2017. doi: 10.1002/rcs.1773. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rcs.1773.

Various. Groundbreaking insights into the implementation of metrics in medical image segmentation. arXiv preprint arXiv:2410.02630, 2024. URL https://arxiv.org/abs/2410.02630. Contains discussion of Normalized Surface Dice.

Y. Zhang et al. Comprehensive clinical usability-oriented contour quality assessment in radiotherapy. *Physica Medica (or similar clinical engineering journal)*, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11711007/. Discusses correlation of APL and Surface Dice with contour efficiency.