Conformal Risk Control for Semantic Uncertainty Quantification in Computed Tomography

Jacopo Teneggi, J. Webster Stayman, Jeremias Sulam

Johns Hopkins University
{jtenegg1,web.stayman,jsulam1}@jhu.edu

Abstract

Uncertainty quantification is necessary for developers, physicians, and regulatory agencies to build trust in machine learning predictors and improve patient care. Beyond measuring uncertainty, it is crucial to express it in clinically meaningful terms that provide actionable insights. This work introduces a conformal risk control (CRC) procedure for organ-dependent uncertainty estimation, ensuring high-probability coverage of the ground-truth image. We first present a high-dimensional CRC procedure that leverages recent ideas of length minimization. We make this procedure semantically adaptive to each patient's anatomy and positioning of organs. Our method, sem-CRC, provides tighter uncertainty intervals with valid coverage on real-world computed tomography data while communicating uncertainty with clinically relevant features.

Introduction

Deep learning predictors are becoming ubiquitous in solving inverse problems in medical imaging, with remarkable performance across diverse modalities and organ systems. Point predictors, however, are limited in their ability to quantify uncertainty, as is often necessary for developers, physicians, and regulatory agencies to verify the safety and reliability of these models in real-world clinical settings. For example, it has been shown that diffusion models can hallucinate the details of a patient's anatomy (Tivnan et al. 2024; Webber and Reader 2024), and robust notions of predictive uncertainty could ameliorate these issues. At the same time, several studies have highlighted the benefits of including uncertainty estimates in computer-aided decision making processes (McCrindle et al. 2021; Faghani et al. 2023; Maruccio et al. 2024; Salvi et al. 2025). This motivates communicating uncertainty in a clinically informed or clinically relevant manner.

Conformal risk control (CRC) (Angelopoulos et al. 2024a) addresses the challenges of measuring the uncertainty of black-box systems without assuming a predictive distribution, having found numerous applications in medicine (Hulsman et al. 2024; Angelopoulos et al. 2024b; Kutiel et al. 2023; Teneggi et al. 2023; Angelopoulos et al. 2022). In imaging, CRC constructs pixel-wise intervals by starting from heuristic notions of uncertainty (e.g., quantile regression (Koenker and Bassett Jr 1978), MC-Dropout (Gal and Ghahramani 2016), or variance of the samples from a diffusion model (Teneggi et al. 2023)), and then conformalizing the resulting sets to achieve risk control. How to minimize interval length in high-dimensional settings is the subject of ongoing research (Kiyani, Pappas, and Hassani 2024; Bars and Humbert 2025; Belhasin et al. 2023; Nehme, Yair, and Michaeli 2023).

In this work, we observe that patients' anatomies vary in size, shape, and positioning of organs, and these variations may unintentionally inflate interval length. We propose to construct *organ-dependent* uncertainty intervals that encompass semantic structures beyond pixels. We achieve this by extending the CRC-equivalent of the K-RCPS procedure (Teneggi et al. 2023), minimizing the mean interval length via convex optimization. Not only does our method, sem-CRC, provide tighter intervals, but it can also guarantee the same level of risk control for each organ rather than cumulatively over a scan. Our work is related to SG-RCPS (Fischer et al. 2024), who study organ-wise risk control in radiotherapy. Here, we focus on computed tomography (CT) data, and our procedure is technically novel. First, we study mean interval length minimization. Second, sem-CRC computes a semantic uncertainty vector whose entries correspond to different organs, whereas SG-RCPS outputs a single scalar that controls all organs-wise risks (Laufer-Goldshtein et al. 2022). We evaluate our method on quantile regression for CT denoising and a simple FBP-UNet reconstruction pipeline using two real-world datasets: TotalSegmentator (Wasserthal et al. 2023) and FLARE23 (Ma et al. 2022). Our contributions apply broadly to any imaging inverse problem and any predictor with a heuristic notion of uncertainty.

Background

Recall that in inverse problems, we aim to retrieve an underlying signal $X \in \mathcal{X}$ from measurements $Y \in \mathcal{Y}$, where $Y = \mathcal{A}(X)$ and the operator $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ cannot be directly inverted (e.g., due to being ill-posed or affected by noise). Herein, we let $\mathcal{X}$ be the space of images with d pixels, i.e. $\mathcal{X} \subseteq [0, 1]^d$.

Quantile regression. A common approach to solving inverse problems is to train a *point predictor* $f : \mathcal{Y} \to \mathcal{X}$ that minimizes a loss function $L(f(y), x)$ over a dataset $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{n}$ of ground-truth signals with their measurements. For example, if L is the squared error then $f(Y) \approx \mathbb{E}[X \mid Y]$. Differently, quantile regression trains a *set predictor* $g : \mathcal{Y} \to 2^{\mathcal{X}}$ such that $\forall j \in [d]$, $g(y)_j = [\hat{q}_\alpha(y)_j, \hat{q}_{1-\alpha}(y)_j]$ where $\hat{q}_t(Y)_j$ is the estimate of the t-level quantile of $\mathbb{P}[X_j \mid Y]$, which can be learned by minimizing the pinball loss (Koenker and Bassett Jr 1978). Thus, quantile regression provides an estimate of uncertainty with intervals length.

Conformal risk control (CRC). The goal of conformal risk control (Angelopoulos et al. 2024a) is to post-process a fixed set predictor g to bound the expectation of its error. More formally, denote $\{g_\lambda\}_{\lambda \in \mathbb{R}_{\geq 0}}$ the family of nested predictors with

$$g_\lambda(y)_j = [\hat{q}_\alpha(y)_j - \lambda, \hat{q}_{1-\alpha}(y)_j + \lambda], \tag{1}$$

and let $\ell(g_\lambda(y), x)$ be any bounded, non-increasing function of λ. Following prior work (Angelopoulos et al. 2022; Teneggi et al. 2023), we will consider the proportion of ground-truth pixels that fall outside of their intervals, i.e.

$$\ell^{01}(g_\lambda(y), x) = \frac{1}{d} \sum_{j \in [d]} \mathbf{1}\{x_j \notin g_\lambda(y)_j\}, \tag{2}$$

which is monotonically non-increasing in $\lambda \in \mathbb{R}_{\geq 0}$ and bounded by 1. Then, for any tolerance $\epsilon > 0$, one can find the smallest parameter $\hat{\lambda}$ that controls the loss in Equation (2). In particular, given a calibration set $S_{\text{cal}} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{n_{\text{cal}}}$, and a test point (X, Y) of exchangeable observations independent of g, the choice of

$$\hat{\lambda} = \inf \left\{ \lambda \in \mathbb{R}_{\geq 0} : \frac{n_{\text{cal}}}{n_{\text{cal}} + 1} \hat{\ell}^{01}_{\text{cal}}(\lambda) + \frac{1}{n_{\text{cal}} + 1} \leq \epsilon \right\} \tag{3}$$

where $\hat{\ell}^{01}_{\text{cal}}(\lambda) = 1/n_{\text{cal}} \sum_{(x,y) \in S_{\text{cal}}} \ell^{01}(g_\lambda(y), x)$ guarantees that

$$\mathbb{E}[\ell^{01}(g_{\hat{\lambda}}(Y), X)] \leq \epsilon, \tag{4}$$

where the expectation is taken over S_{cal} and (X, Y).

High-dimensional risk control. As noted by (Teneggi et al. 2023), using the same scalar λ for all pixels inflates the mean interval length of the conformalized sets. To overcome this limitation, they propose to assign each pixel to one of K groups with some shared statistics. More precisely, they consider a partition matrix $M \in \{0, 1\}^{d \times K}$, and use a vector-valued parameter $\boldsymbol{\lambda}_K = [\lambda_1, \ldots, \lambda_K] \in \mathbb{R}^K_{\geq 0}$ such that $\boldsymbol{\lambda} = M\boldsymbol{\lambda}_K \in \mathbb{R}^d_{\geq 0}$ and

$$g_{\boldsymbol{\lambda}}(y)_j = [\hat{q}_\alpha(y)_j - \lambda_j, \hat{q}_{1-\alpha}(y)_j + \lambda_j]. \tag{5}$$

Then, for a fixed anchor point $\widetilde{\boldsymbol{\lambda}}_K \in \mathbb{R}^K_{\geq 0}$, choosing

$$\hat{\boldsymbol{\lambda}} = \inf \left\{ \boldsymbol{\lambda} \in M\widetilde{\boldsymbol{\lambda}}_K + \omega \mathbf{1}_d,\ \omega \in \mathbb{R} :\ R^+_{\text{cal}}(\boldsymbol{\lambda}) \leq \epsilon \right\}, \tag{6}$$

where

$$R^+_{\text{cal}}(\boldsymbol{\lambda}) = \frac{n_{\text{cal}}}{n_{\text{cal}} + 1} \hat{\ell}^{01}_{\text{cal}}(\boldsymbol{\lambda}) + \frac{1}{n_{\text{cal}} + 1} \tag{7}$$

controls risk as in Equation (4). Note that (Teneggi et al. 2023) introduced their method for risk controlling prediction sets (RCPSs) (Bates et al. 2021), but it applies to CRC as well. The anchor $\widetilde{\boldsymbol{\lambda}}_K \in \mathbb{R}^K_{\geq 0}$ is arbitrary, but it should be chosen to minimize the mean interval length. The proposed method, K-CRC, introduces ℓ^γ for $\gamma \in (0, 1)$: a convex upper-bound to ℓ^{01}. Then, it solves the following optimization problem

$$\begin{aligned} \widetilde{\boldsymbol{\lambda}}_K = &\arg\min_{\boldsymbol{\lambda}_K \in \mathbb{R}^K_{\geq 0}} \sum_{k \in [K]} n_k \lambda_k \\ &\text{s.t.} \quad \hat{\ell}^\gamma_{\text{opt}}(M\boldsymbol{\lambda}_K) \leq \epsilon, \end{aligned} \tag{PK}$$

where n_k is the number of pixels in group k. We stress that in this procedure, the calibration set S_{cal} needs to be split in S_{opt} and $\widetilde{S}_{\text{cal}}$, such that the former is used to solve (PK) and the latter to find $\hat{\boldsymbol{\lambda}}$ as in Equation (6).

With this background, we now present the main contributions of our work.

Semantic Uncertainty Quantification

Observe that the partition matrix M that assigns each of the d pixels to one of K groups does not depend on the measurement y. This choice is effective when the semantic content of each pixel is similar across observations (e.g., face images can be aligned and centered). However, CT data is heterogeneous, and a fixed partition matrix may unnecessarily increase the mean interval length.

In this work, we leverage foundational segmentation models (Qu et al. 2023; Li, Yuille, and Zhou 2024) to construct organ-dependent uncertainty intervals. Our method, sem-CRC, extends K-CRC to instance-dependent memberships $s(y) \in [K]^d$. This decouples optimizing the mean interval length from the pixel domain, and it reflects the uncertainty of the model in terms of semantic—and clinically meaningful—structures. Formally, let $s : \mathcal{Y} \to [K]^d$ be a fixed segmentation model such that, for a vector $\boldsymbol{\lambda}_{\text{sem}} \in \mathbb{R}^K_{\geq 0}$, the family of nested set predictors $\{g_{\boldsymbol{\lambda}_{\text{sem}}}\}$ is given by

$$g_{\boldsymbol{\lambda}_{\text{sem}}}(y)_j = [\hat{q}_\alpha(y)_j - \lambda_{s(y)_j}, \hat{q}_{1-\alpha}(y)_j + \lambda_{s(y)_j}]. \tag{8}$$

Note that, differently from $g_{\boldsymbol{\lambda}}(y)$ in Equation (5), the same pixel j may receive different assignments in different scans depending on the measurement y. Our work does not study the performance of s, and calibration of segmentation models is subject of complementary research (Mossina, Dalmau, and Andéol 2024; Wundram et al. 2024; Davenport 2024; Brunekreef et al. 2024). We will proceed analogously to the above, i.e. finding an anchor $\widetilde{\boldsymbol{\lambda}}_{\text{sem}}$ that minimizes the mean interval length $\bar{I}_{\boldsymbol{\lambda}_{\text{sem}}}(y)$, and then backtracking along the line $\widetilde{\boldsymbol{\lambda}}_{\text{sem}} + \omega \mathbf{1}_K$ to control risk. Start by noting that the mean interval length can be expressed as

$$\bar{I}_{\boldsymbol{\lambda}_{\text{sem}}}(y) = \frac{1}{d} \sum_{j \in [d]} I(y)_j + \frac{1}{d} \sum_{k \in [K]} |\mathcal{S}_k(y)| \lambda_k, \tag{9}$$

where $I(y)_j = \hat{q}_{1-\alpha}(y)_j - \hat{q}_\alpha(y)_j$ is the width of the prediction interval for pixel j, and $\mathcal{S}_k(y) = \{j \in [d] : s(y)_j = k\}$

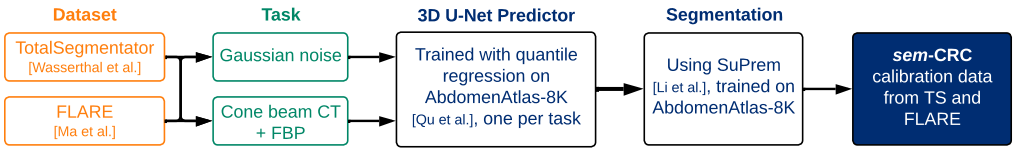

Figure 1: Illustration of our experimental setup.

is the set of voxels that belong to organ k for observation y. We can see that the mean interval length is still a function of the sum of the λ_k's, but the multiplicative factors now depend on y as well. So, it becomes necessary to minimize the mean interval length in expectation over Y. We extend the original optimization problem (PK) to its semantic version

$$\widetilde{\boldsymbol{\lambda}}_{\text{sem}} = \underset{\boldsymbol{\lambda}_{\text{sem}} \in \mathbb{R}^K_{\geq 0}}{\arg\min} \sum_{k \in [K]} \mathbb{E}_Y[|\mathcal{S}_k(Y)|]\lambda_k \quad \text{s.t.} \quad \hat{\ell}^{\gamma}_{\text{opt}}(\boldsymbol{\lambda}_{\text{sem}}) \leq \epsilon, \tag{Psem}$$

where, in practice, we estimate the objective over S_{opt}. To conclude, we choose

$$\hat{\boldsymbol{\lambda}}_{\text{sem}} = \inf\left\{\boldsymbol{\lambda}_{\text{sem}} \in \widetilde{\boldsymbol{\lambda}}_{\text{sem}} + \omega \mathbf{1}_K : R^+_{\text{cal}}(\boldsymbol{\lambda}_{\text{sem}}) \leq \epsilon\right\}, \tag{10}$$

and we state the validity of sem-CRC in the following proposition.[1]

Proposition 1. *For a risk tolerance $\epsilon > 0$, segmentation model $s : \mathcal{Y} \to [K]^d$, anchor point $\widetilde{\boldsymbol{\lambda}}_{sem} \in \mathbb{R}^K_{\geq 0}$, and exchangeable calibration and test points $S_{cal} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{n_{cal}}$, (X, Y), the choice of $\hat{\boldsymbol{\lambda}}_{sem}$ as in Equation* (10) *provides risk control, i.e.*

$$\mathbb{E}[\ell^{01}(g_{\hat{\boldsymbol{\lambda}}_{sem}}(Y), X)] \leq \epsilon. \tag{11}$$

Proof. Let $\boldsymbol{\lambda}_{\text{sem}}(\omega) = \widetilde{\boldsymbol{\lambda}}_{\text{sem}} + \omega \mathbf{1}_K$, $\omega \in \mathbb{R}$, and note that $\ell^{01}(g_{\boldsymbol{\lambda}_{\text{sem}}(\omega)}(y), x)$ is bounded by 1 and monotonically non-increasing in ω. Since s is fixed, the random functions $L_i(\omega) = \ell^{01}(g_{\boldsymbol{\lambda}_{\text{sem}}(\omega)}(Y^{(i)}), X^{(i)})$ and $L(\omega) = \ell^{01}(g_{\boldsymbol{\lambda}_{\text{sem}}(\omega)}(Y), X)$ are exchangeable. The result then follows by applying (Angelopoulos et al. 2024a, Theorem 1) to ω. □

We remark that sem-CRC also relies on splitting the calibration set S_{cal} into S_{opt} to solve (Psem), and $\widetilde{S}_{\text{cal}}$ to find $\hat{\boldsymbol{\lambda}}_{\text{sem}}$ as in Equation (10). Furthermore, and naturally, the method requires performing inference with the same segmentation model used for calibration. We regard semantic calibration with respects to ground-truth segmentations as an extension of this work.

Controlling risk for each organ. Clinical tasks may require different organs to have the same level of reconstruction accuracy, but $\hat{\boldsymbol{\lambda}}_{\text{sem}}$ may overcover easy-to-reconstruct ones while undercovering others. For example, as noted by (Fischer et al. 2024), per-organ coverage is critical to avoid treatment errors for tumor resection planning or organ transplant evaluation, at the cost of larger uncertainty intervals. On the other hand, interval length minimization across several organs informs on the distribution of the error of the model for tasks such as whole-abdomen segmentation or total lesion volume measurement. Thus, we specialize sem-CRC to control risk with the same tolerance ϵ for each segmented structure, and we call this variation $\overline{sem}$-CRC. With this, our contributions allow clinicians to use the minimal dose that guarantees risk control for a target organ with uncertainty intervals shorter than a task-driven tolerance, and regulatory agencies to potentially issues standards accordingly. Denote

$$\ell^{01}_k(g_{\boldsymbol{\lambda}_{\text{sem}}}(y), x) = \frac{1}{|\mathcal{S}_k(y)|} \sum_{j \in \mathcal{S}_k(y)} \mathbf{1}\{x_j \notin g_{\boldsymbol{\lambda}_{\text{sem}}}(y)_j\} \tag{12}$$

the proportion of pixels in organ k (e.g., liver) that fall outside of their intervals, and let $\boldsymbol{e}_k$ be the k^{th} standard basis vector. The choice of $\hat{\boldsymbol{\lambda}}_{\overline{\text{sem}}} \in \mathbb{R}^K_{\geq 0}$ with

$$\hat{\lambda}_{\overline{\text{sem}},k} = \inf\left\{\lambda \in \mathbb{R}_{\geq 0} : R^+_{k,\text{cal}}(\widetilde{\boldsymbol{\lambda}}_{\text{sem}} + \lambda \boldsymbol{e}_k) \leq \epsilon\right\}, \tag{13}$$

where

$$R^+_{k,\text{cal}}(\boldsymbol{\lambda}) = \frac{n_{\text{cal}}}{n_{\text{cal}}+1}\hat{\ell}^{01}_{k,\text{cal}}(\boldsymbol{\lambda}) + \frac{1}{n_{\text{cal}}+1} \tag{14}$$

provides risk control for each organ, that is $\mathbb{E}[\ell^{01}_k(g_{\hat{\boldsymbol{\lambda}}_{\overline{\text{sem}}}}(Y), X)] \leq \epsilon$, $k = 1, \ldots, K$. This follows by applying Proposition 1 to each dimension of $\hat{\boldsymbol{\lambda}}_{\overline{\text{sem}}}$. We briefly remark this is different from multiple risk control with one scalar λ as in (Angelopoulos et al. 2024a; Laufer-Goldshtein et al. 2022; Fischer et al. 2024), and that the equivalent for RCPS requires multiple hypothesis testing correction.

Experiments

We compare CRC, K-CRC, and sem-CRC for denoising and for a basic FBP-UNet reconstruction task on TotalSegmentator (Wasserthal et al. 2023) (1,429 scans) and on the first 1,000 scans from the training split of the FLARE23 (Ma et al. 2022) challenge (see Figure 1 for an illustration of the experimental setup). We resample the FLARE23 dataset at $1.5\,\text{mm} \times 1.5\,\text{mm} \times 3.0\,\text{mm}$ resolution, and we window all scans between -175 HU and 250 HU. For denoising, we add independent Gaussian noise with $\sigma = 0.2$; for reconstruction, we use the ODL library (Adler, Kohr, and Öktem 2017) with ASTRA (Van Aarle et al. 2016, 2015) to simulate a helical cone beam geometry. We set the pitch adaptively to

[1] We present results for CRC, but our method generalizes to RCPSs as well.

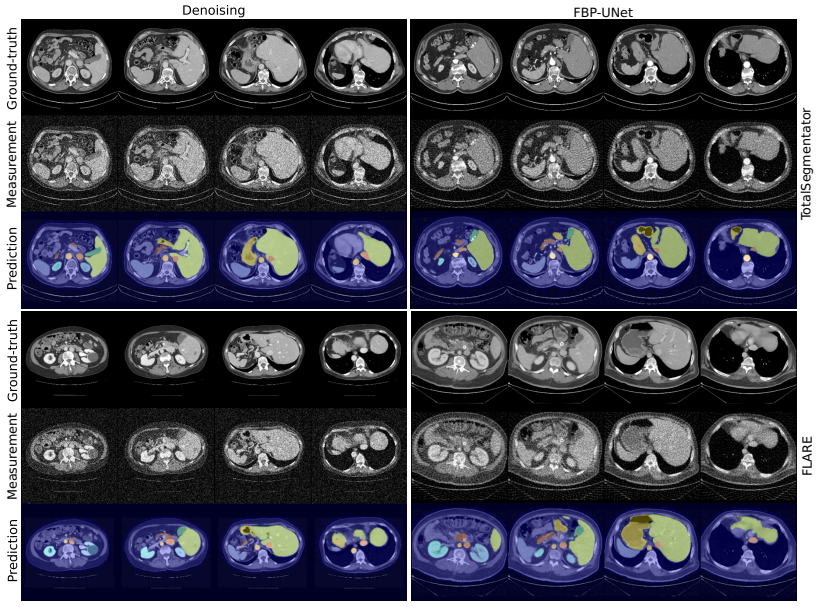

Figure 2: Example calibration data: ground-truth, measurement, and segmented predictions for both tasks and datasets.

		TotalSegmentator		FLARE23	
Task	Procedure	Risk	Length ($\times 10^{-2}$)	Risk	Length ($\times 10^{-2}$)
Denoising	CRC	0.095 ± 0.006	11.60 ± 0.21	0.096 ± 0.004	9.16 ± 0.09
	K-CRC	0.097 ± 0.006	9.37 ± 0.20	0.096 ± 0.006	6.81 ± 0.21
	sem-CRC	0.098 ± 0.006	$\mathbf{8.72 \pm 0.18}$	0.095 ± 0.006	$\mathbf{6.36 \pm 0.11}$
	$\overline{sem}$-CRC	0.055 ± 0.004	11.84 ± 0.20	0.056 ± 0.003	8.06 ± 0.16
FBP-UNet	CRC	0.098 ± 0.007	10.43 ± 0.23	0.095 ± 0.006	6.19 ± 0.09
	K-CRC	0.098 ± 0.009	9.32 ± 0.13	0.095 ± 0.003	6.20 ± 0.14
	sem-CRC	0.097 ± 0.007	$\mathbf{8.95 \pm 0.19}$	0.095 ± 0.006	$\mathbf{6.18 \pm 0.13}$
	$\overline{sem}$-CRC	0.059 ± 0.005	12.43 ± 0.20	0.057 ± 0.003	7.72 ± 0.17

Table 1: Summary of calibration results as mean and standard deviation over 20 independent runs of each calibration procedure with risk tolerance $\epsilon = 0.10$.

cover the entire volume in 8 turns, and acquire data over 1,000 angles with a detector of size 512 pixels $\times$ 128 pixels. We model low-dose measurement as linear Poisson noise with $I_0 = 1,000$. We chose these settings to highlight our method's performance on a challenging task. For each task, we use MONAI (Cardoso et al. 2022) to train a 3D UNet (Ronneberger, Fischer, and Brox 2015) (≈ 5 M parameters, ROI of 96^3 voxels) with quantile regression ($\alpha = 0.1$, i.e. the 10^{th} and 90^{th} quantiles) on the AbdomenAtlas-8K dataset (Qu et al. 2023) ($5,195$ scans).

We segment 9 structures: spleen, kidneys, gallbladder, liver, stomach, aorta, inferior vena cava (IVC), and pancreas using SuPrem (Li, Yuille, and Zhou 2024), a state-of-the-art general-purpose segmentation model for medical imaging. All remaining voxels that are not background are labeled generically as "body". We remark that the segmentation model s is introduced as a function of the measurement for the sake of generality. Here, we segment the predictions of the 3D UNet, but any strategy independent of the calibration data would be valid. To solve (Psem) over a distribution of volumes that represents all organs, we select 4 equidistant slices from the window of 48 that maximizes the segmentation volume. Finally, we center-crop or pad slices to 256 voxels $\times$ 256 voxels for calibration.

Since sem-CRC relies on a fixed segmentation model, we evaluate predictions in terms of mean structure-wise F1 score between the segmented outputs and the ground-truth annotations over 200 random volumes. For the TotalSegmentator dataset, we obtain 0.85 ± 0.07 and 0.83 ± 0.08 for denoising and FBP-UNet, respectively; and, equivalently, 0.88 ± 0.06 and 0.87 ± 0.07 for the FLARE23 dataset. Although we see a slight drop in performance compared to the metrics reported in (Li, Yuille, and Zhou 2024), these results confirm predictions are of reasonable quality for segmentation, and we include some examples in Figure 2.

We set the error tolerance to $\epsilon = 0.10$, allowing at most 10% of ground-truth voxels to fall outside their prediction intervals. Each calibration procedure is run 20 times on in-

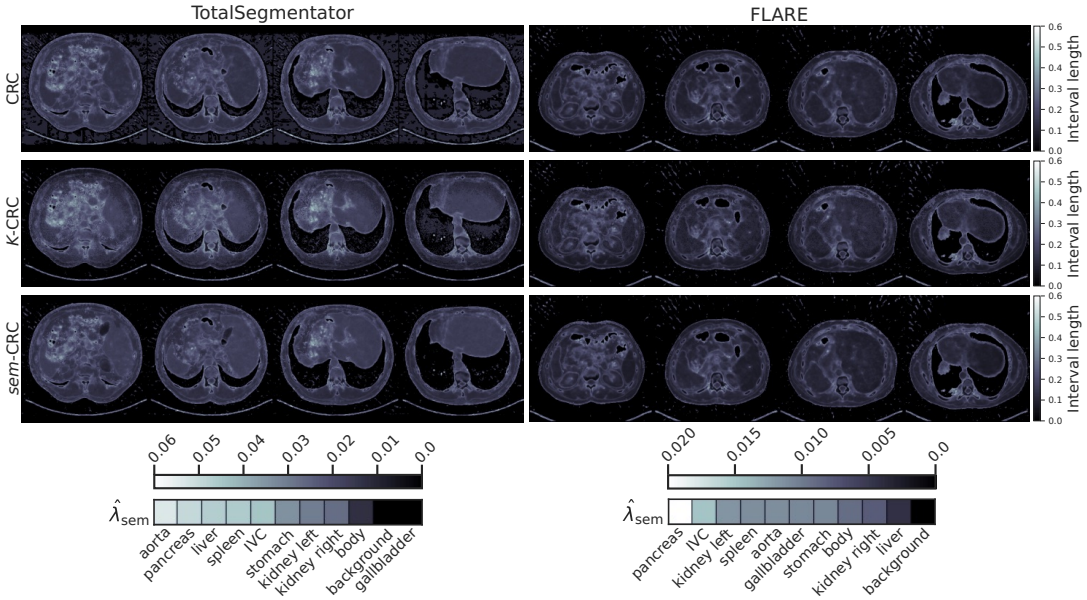

Figure 3: Example conformalized uncertainty maps on one volume per dataset with each calibration method for the FBP-UNet pipeline. The bottom row shows $\hat{\boldsymbol{\lambda}}_{\text{sem}}$, the semantic uncertainty parameter learned by our method, sem-CRC.

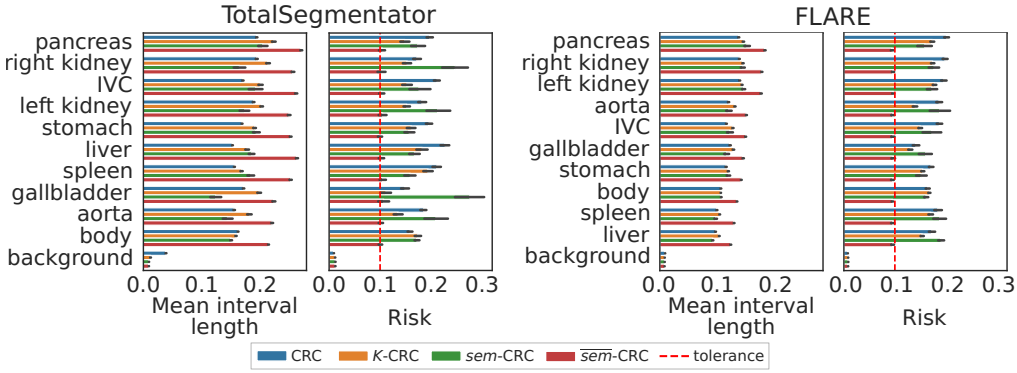

Figure 4: Mean interval length and risk stratified by organ for the FBP-UNet task across all calibration procedures and datasets. $\overline{sem}$-CRC is the only procedure that guarantees risk control for each organ.

dependent subsets of $n_{\text{cal}} = 512$ scans, with risk estimated on $n_{\text{test}} = 128$ scans. We allocate $n_{\text{opt}} = 32$ calibration samples to solve (PK) and (Psem), ensuring a fair comparison across methods. For K-CRC, we follow (Teneggi et al. 2023) and construct the assignment matrix M by grouping voxels into $K = 4$ quantiles of the loss on the optimization set. Finally, to solve (PK) efficiently, we subsample $d_{\text{opt}} = 50$ voxels (much smaller than 256^2) stratified by membership; and for (Psem), we ensure the smallest organ has a support of at least $d_{\min} = 2$ voxels by subsampling $d_{\text{opt}} = d_{\min} / \min_k \mathbb{E}[|\mathcal{S}_k|]$ dimensions ($d_{\text{opt}} \approx 3,000$). Solving subsampled problems reduces complexity to the order of seconds.

Table 1 summarizes risk and mean interval length across all datasets and tasks. All procedures are *valid*, i.e. they control risk at level ϵ. Our method, sem-CRC, consistently provides the shortest uncertainty intervals. On the other hand, and as expected, controlling risk for each organ with $\overline{sem}$-CRC increases the mean interval length. Figure 3 compares the conformalized uncertainty maps obtained with each method on the same volume, and it includes the vector $\hat{\boldsymbol{\lambda}}_{\text{sem}}$ learned by sem-CRC. The uncertainty maps generated by sem-CRC are sharper and contain fewer artifacts thanks to using instance-level information. Furthermore, $\hat{\boldsymbol{\lambda}}_{\text{sem}}$ directly informs on which organs have higher levels of uncertainty, depicting how the same model may display different uncertainty patterns across different populations. These findings are fundamental to the responsible use of general-purpose machine learning models across centers serving diverse demographics. Finally, Figure 4 highlights the difference between controlling risk for each organ or cumulatively over a volume: all methods but $\overline{sem}$-CRC achieve risk control by overcovering background and undercovering organs. Our methodology gives users the flexibility to specify which

organs they desire to control risk for depending on the clinical task at hand.

Conclusions

Modern deep learning models are widely used for image reconstruction, including computed tomography. However, they often provide only point-wise estimates, lacking statistically valid uncertainty measures. This work proposes a conformal prediction approach that generates uncertainty intervals with controlled risk at any user-specified level. By integrating high-dimensional calibration and state-of-the-art segmentation models, our method, sem-CRC, produces organ-dependent uncertainty sets that are adaptive to each patient. Moreover, it can control risk *for each organ*. Not only does sem-CRC provide the tightest uncertainty set, but also it communicates findings with clinically meaningful anatomical structures.

Acknowledgments

This research was supported by NSF CAREER Award CCF 2239787 and by NIH R01CA287422.

References

Adler, J.; Kohr, H.; and Öktem, O. 2017. Operator Discretization Library (ODL).

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2024a. Conformal Risk Control. In *The Twelfth International Conference on Learning Representations*.

Angelopoulos, A. N.; Kohli, A. P.; Bates, S.; Jordan, M.; Malik, J.; Alshaabi, T.; Upadhyayula, S.; and Romano, Y. 2022. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, 717–730. PMLR.

Angelopoulos, A. N.; Pomerantz, S.; Do, S.; Bates, S.; Bridge, C. P.; Elton, D. C.; Lev, M. H.; González, R. G.; Jordan, M. I.; and Malik, J. 2024b. Conformal Triage for Medical Imaging AI Deployment. *medRxiv*, 2024–02.

Bars, B. L.; and Humbert, P. 2025. On Volume Minimization in Conformal Regression. *arXiv preprint arXiv:2502.09985*.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6): 1–34.

Belhasin, O.; Romano, Y.; Freedman, D.; Rivlin, E.; and Elad, M. 2023. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3321–3333.

Brunekreef, J.; Marcus, E.; Sheombarsing, R.; Sonke, J.-J.; and Teuwen, J. 2024. Kandinsky conformal prediction: efficient calibration of image segmentation algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4135–4143.

Cardoso, M. J.; Li, W.; Brown, R.; Ma, N.; Kerfoot, E.; Wang, Y.; Murrey, B.; Myronenko, A.; Zhao, C.; Yang, D.; et al. 2022. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*.

Davenport, S. 2024. Conformal confidence sets for biomedical image segmentation. *arXiv preprint arXiv:2410.03406*.

Faghani, S.; Moassefi, M.; Rouzrokh, P.; Khosravi, B.; Baffour, F. I.; Ringler, M. D.; and Erickson, B. J. 2023. Quantifying uncertainty in deep learning of radiologic images. *Radiology*, 308(2): e222217.

Fischer, P.; Willms, H.; Schneider, M.; Thorwarth, D.; Muehlebach, M.; and Baumgartner, C. F. 2024. Subgroup-Specific Risk-Controlled Dose Estimation in Radiotherapy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 696–706. Springer.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Hulsman, R.; Comte, V.; Bertolini, L.; Wiesenthal, T.; Gallardo, A. P.; and Ceresa, M. 2024. Conformal Risk Control for Pulmonary Nodule Detection. *arXiv preprint arXiv:2412.20167*.

Kiyani, S.; Pappas, G.; and Hassani, H. 2024. Length optimization in conformal prediction. *arXiv preprint arXiv:2406.18814*.

Koenker, R.; and Bassett Jr, G. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Kutiel, G.; Cohen, R.; Elad, M.; Freedman, D.; and Rivlin, E. 2023. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *International Workshop on Trustworthy Machine Learning for Healthcare*, 163–176. Springer.

Laufer-Goldshtein, B.; Fisch, A.; Barzilay, R.; and Jaakkola, T. 2022. Efficiently controlling multiple risks with pareto testing. *arXiv preprint arXiv:2210.07913*.

Li, W.; Yuille, A.; and Zhou, Z. 2024. How well do supervised models transfer to 3d image segmentation. In *The Twelfth International Conference on Learning Representations*, volume 1.

Ma, J.; Zhang, Y.; Gu, S.; An, X.; Wang, Z.; Ge, C.; Wang, C.; Zhang, F.; Wang, Y.; Xu, Y.; et al. 2022. Fast and low-GPU-memory abdomen CT organ segmentation: the flare challenge. *Medical Image Analysis*, 82: 102616.

Maruccio, F. C.; Eppinga, W.; Laves, M.-H.; Navarro, R. F.; Salvi, M.; Molinari, F.; and Papaconstadopoulos, P. 2024. Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation. *Physics in Medicine & Biology*, 69(3): 035007.

McCrindle, B.; Zukotynski, K.; Doyle, T. E.; and Noseworthy, M. D. 2021. A radiology-focused review of predictive uncertainty for AI interpretability in computer-assisted segmentation. *Radiology: Artificial Intelligence*, 3(6): e210031.

Mossina, L.; Dalmau, J.; and Andéol, L. 2024. Conformal Semantic Image Segmentation: Post-hoc Quantification of Predictive Uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3574–3584.

Nehme, E.; Yair, O.; and Michaeli, T. 2023. Uncertainty quantification via neural posterior principal components.

Advances in Neural Information Processing Systems, 36: 37128–37141.

Qu, C.; Zhang, T.; Qiao, H.; Tang, Y.; Yuille, A. L.; Zhou, Z.; et al. 2023. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36: 36620–36636.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Salvi, M.; Seoni, S.; Campagner, A.; Gertych, A.; Acharya, U. R.; Molinari, F.; and Cabitza, F. 2025. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, 105846.

Teneggi, J.; Tivnan, M.; Stayman, W.; and Sulam, J. 2023. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, 33940–33960. PMLR.

Tivnan, M.; Yoon, S.; Chen, Z.; Li, X.; Wu, D.; and Li, Q. 2024. Hallucination Index: An Image Quality Metric for Generative Reconstruction Models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 449–458. Springer.

Van Aarle, W.; Palenstijn, W. J.; Cant, J.; Janssens, E.; Bleichrodt, F.; Dabravolski, A.; De Beenhouwer, J.; Joost Batenburg, K.; and Sijbers, J. 2016. Fast and flexible X-ray tomography using the ASTRA toolbox. *Optics express*, 24(22): 25129–25147.

Van Aarle, W.; Palenstijn, W. J.; De Beenhouwer, J.; Altantzis, T.; Bals, S.; Batenburg, K. J.; and Sijbers, J. 2015. The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*, 157: 35–47.

Wasserthal, J.; Breit, H.-C.; Meyer, M. T.; Pradella, M.; Hinck, D.; Sauter, A. W.; Heye, T.; Boll, D. T.; Cyriac, J.; Yang, S.; et al. 2023. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5).

Webber, G.; and Reader, A. J. 2024. Diffusion models for medical image reconstruction. *BJR— Artificial Intelligence*, 1(1): ubae013.

Wundram, A. M.; Fischer, P.; Mühlebach, M.; Koch, L. M.; and Baumgartner, C. F. 2024. Conformal performance range prediction for segmentation output quality control. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, 81–91. Springer.