

CARE: CONFIDENCE-AWARE RATIO ESTIMATION FOR MEDICAL BIOMARKERS

Anonymous authors

Paper under double-blind review

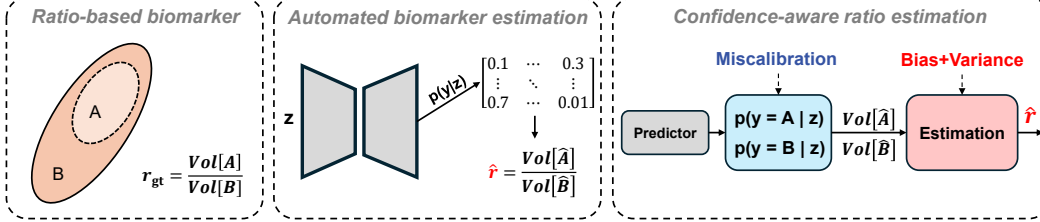


Figure 1: **Overview of CARE.** In automated medical imaging analysis, biomarkers are often computed from network predictions. To quantify the uncertainty of ratio-based biomarkers, we introduce CARE, a confidence-aware estimation method that provides reliable confidence intervals.

ABSTRACT

Ratio-based biomarkers – such as the proportion of necrotic tissue within a tumor – are widely used in clinical practice to support diagnosis, prognosis, and treatment planning. These biomarkers are typically estimated from soft segmentation outputs by computing region-wise ratios. Despite the high-stakes nature of clinical decision making, existing methods provide only point estimates, offering no measure of uncertainty. In this work, we propose a unified *confidence-aware* framework for estimating ratio-based biomarkers. Our uncertainty analysis stems from two observations: i) the probability ratio estimator inherently admits a statistical confidence interval regarding local randomness (bias and variance), ii) the segmentation network is not perfectly calibrated. We conduct a systematic analysis of error propagation in the segmentation-to-biomarker pipeline and identify model miscalibration as the dominant source of uncertainty. We leverage tunable parameters to control the confidence level of the derived bounds, allowing adaptation towards clinical practice. Extensive experiments show that our method produces statistically sound confidence intervals, with tunable confidence levels, enabling more trustworthy application of predictive biomarkers in clinical workflows.

1 INTRODUCTION

The success of deep learning in medical image analysis, particularly since the introduction of UNet architectures (Ronneberger et al., 2015; Isensee et al., 2021), has enabled automated segmentation of anatomical and pathological structures across a range of clinical imaging tasks. However, segmentation is rarely the end goal in clinical workflows. Instead, it often serves as an intermediate step toward computing biomarkers – quantitative metrics such as volumes (Popordanoska et al., 2021; Rousseau et al., 2025; Kazerooni et al., 2023; Abdusalomov et al., 2023) and fraction scores (Ronneberger et al., 2015; Isensee et al., 2021; Bahna et al., 2022; Kim et al., 2008) – that are used to assess disease progression, guide treatment decisions, or monitor therapeutic response. As shown in Fig. 1, the ratio-based biomarker is usually derived from two volume measurements. Since segmentation models provide per-pixel prediction, they allow automated ratio estimation. Nevertheless, relying solely on a single point estimate offers no quantification of uncertainty, which limits the clinical adoption and undermines its value as a decision-making reference. To address this, we study confidence-aware ratio estimation for medical biomarkers.

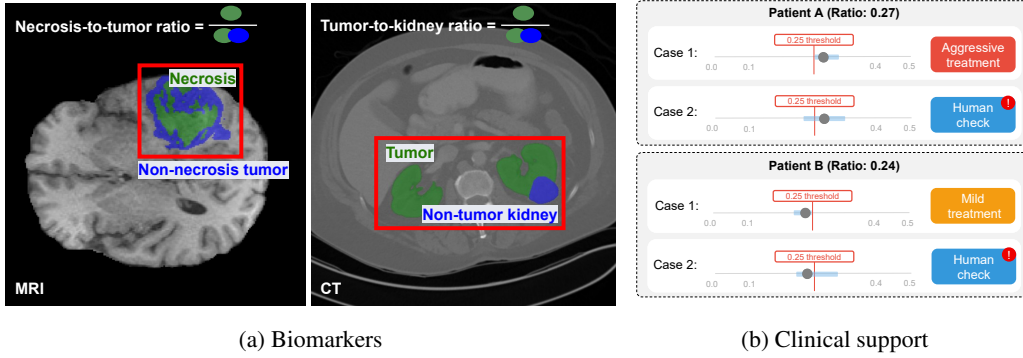


Figure 2: **Medical background of ratio estimation and its role in clinical support.** (a): Ratio-based biomarkers (Baid et al., 2021; Myronenko et al., 2023) exist in many organs and modalities. (b): An illustrative example where a high-risk threshold is defined as 0.25; CARE calls for human check when the confidence interval crosses the threshold.

As shown in Fig. 2a, ratio-based biomarkers are widely utilized across various organs and imaging modalities. These biomarkers provide quantitative measures for clinical decision-making. For example, the necrosis-to-tumor ratio (NTR) (Henker et al., 2019; 2017) quantifies the proportion of necrotic (non-viable) tissue within a tumor, while the tumor-to-kidney ratio (TKR) Herts et al. (2002) indicates the extent of tumor infiltration within the kidney. Accurate estimation of these biomarkers is crucial for supporting personalized treatments and monitoring their efficacy. A straightforward method for computing these ratios involves using segmentation models to identify the subregion and the whole foreground region, and then calculating the ratio based on averaged softmax confidence scores over these regions. However, the interpretation of this point estimate can change once the confidence interval is considered. As illustrated in Fig. 2b, consider a clinical threshold of 0.25 for initiating aggressive treatment. Based on point estimates alone, Patient A would receive aggressive treatment (high ratio) while Patient B would receive mild treatment (low ratio). However, if the associated confidence interval spans the decision threshold (case 2), the estimation is flagged for mandatory expert review to mitigate potential misdiagnosis risk. Such double-check procedures are essential in clinical practice, as they provide an additional safeguard for patients and enhance the robustness of downstream decision-making. Despite their clinical importance, most efforts still focus on improving upstream segmentation accuracy (Ronneberger et al., 2015; Isensee et al., 2021; Hatamizadeh et al., 2021), while the uncertainty and reliability of downstream ratio-based biomarkers remain largely unexplored.

We propose CARE, the *first confidence-aware estimation framework specifically for ratio-based biomarkers*, offering several key advantages: i) **guaranteed coverage**, *i.e.*, the actual coverage probability of containing the true ratio is greater than the stated nominal confidence level; ii) **instance-wise adaptiveness**, *i.e.*, providing dynamic intervals that capture varying uncertainty degrees; iii) **tunable** confidence level with user-controlled tightness; iv) applicable as a **plug-in** module to any pretrained NN requiring neither architectural modifications nor training from scratch; v) computationally **efficient**, avoiding multiple sampling or repeated forward passes.

Furthermore, we identify sources of error and quantify their individual impacts on the overall confidence intervals. Specifically, we establish a ratio estimator bound using Markov’s inequality (Resnick, 2003) and derive a squared error estimator from volume predictions. We also address the issue of overconfident predictions from deep learning models, which represents another critical limitation undermining ratio estimation reliability. To quantify the error caused by miscalibration, we provide theoretical insights into the relationship between model calibration and ratio estimation and propose a miscalibration-based bound, building on recent advances in calibration error (CE) estimation (Guo et al., 2017; Popordanoska et al., 2022).

Beyond theoretical contributions, our framework is designed for practical clinical deployment. CARE operates with linear time complexity $\mathcal{O}(n)$, making it lightweight and reliable. Compared with Bayesian methods, CARE is well-suited for clinical settings where real-time performance is essential and large-scale computing resources are unavailable. Experiments further confirm that the proposed confidence bounds are conservative, adaptive and computationally efficient.

In summary, our main **contributions** are:

- ① We propose CARE, a principled framework for confidence-aware ratio estimation for medical biomarkers in an automated estimation workflow (Sec. 3).
- ② We analyze the sources of error across the entire segmentation-to-biomarker pipeline (Sec. 3) and empirically demonstrate that miscalibration is the dominant factor (Sec. 4).
- ③ Experiments confirm that CARE effectively tracks the prediction uncertainty, represented as the coverage of erroneous predictions and the distinguishability of segmentation difficulties (Sec. 4).

2 RELATED WORK

Ratio-based biomarkers are quantitative metrics that express the relative size, volume, or intensity of a target anatomical structure as a proportion of a reference region (Fig. 1). They are widely used across clinical domains to capture compositional, structural and functional changes, enabling standardized assessment of disease progression and treatment response. Examples include: ejection fraction – representing the fraction of blood ejected from the ventricle during each cardiac cycle; coronary artery stenosis – quantifying the percent narrowing of a coronary vessel, and fat fraction – measuring the proportion of fat within an organ such as liver or kidney. Ratio-based biomarkers are particularly valuable for detailed tumor characterization (Fig. 2). Key metrics include necrosis-to-tumor ratio (NTR) and core-to-tumor ratio (CTR), which quantify the internal structure of the tumor, as well as tumor invasion rate, which reflects the extent of tumor infiltration into surrounding tissues. In summary, the ratio-based measures offer standardized, comparable metrics that can be applied across imaging modalities, organs, and disease contexts.

Typically, clinicians compute these ratios using volumetric information from imaging data (*e.g.*, MRI) (Henker et al., 2019; 2017). With the advancement of computational pathology and the growing availability of annotated medical data, recent studies (Ye et al., 2023) have developed AI-based workflows for automated ratio assessment. These methods offer scalable and consistent evaluations, effectively overcoming the limitations of subjective human judgment in manual assessments. Despite promising developments, existing methods typically provide only point estimates (Ho et al., 2020), neglecting the associated uncertainty. Although intuitive, results computed from the outputs of segmentation networks inherit the known overconfidence tendency of neural networks (Guo et al., 2017). As a result, naïve ratio estimations from miscalibrated outputs are often biased from true values. Current research predominantly focuses on improving network calibration and segmentation accuracy (Rousseau et al., 2025; Wang et al., 2023; Mehrtash et al., 2020; Wang et al., 2022; Hatamizadeh et al., 2021), while overlooking the downstream task of biomarker estimation. Our work addresses this gap by proposing a confidence-aware framework for ratio estimation from segmentation models.

Uncertainty quantification (UQ) provides many statistical methods to estimate prediction uncertainty. *Conformal prediction (CP)* (Vovk et al., 1999; Papadopoulos et al., 2002; Vovk et al., 2005; Angelopoulos & Bates, 2021; Karimi & Samavi, 2023; Angelopoulos & Bates, 2021) constructs prediction intervals that guarantee valid coverage under finite samples, without any distributional assumptions. Its key strength is the distribution-free nature and finite-sample validity, providing strong theoretical guarantees regardless of the base predictive model. *Resampling methods* are non-parametric techniques for estimating the sampling distribution of a statistic, applicable when the underlying distribution is unknown or difficult to derive. Specifically, *Bootstrapping* (Mooney et al., 1993; Freedman, 1981) repeatedly samples N data points with replacement from the original data, whereas *subsampling* (Politis & Romano, 1994) takes a subset of the original data without replacement, repeating the process multiple times to construct an empirical distribution of the statistic. *Bayesian methods* achieve robust segmentation by averaging multiple predictions, using techniques like deep ensemble (Lakshminarayanan et al., 2017) and Monte Carlo dropout (Srivastava et al., 2014). These approaches enable confidence interval estimation by computing standard deviation across several feedforward inferences. However, they require proper prior specification and cannot provide tunable quantiles due to the limited number of inference samples (usually ≤ 10). Moreover, these universal methods are either computationally expensive or fail to provide informative conclusions.

Calibration error (CE) estimation has attracted extensive research attention (Kull & Flach, 2015; Vaicenavicius et al., 2019; Kumar et al., 2019; Zhang et al., 2020; Popordanoska et al., 2022; Gruber & Buettner, 2022). In medical segmentation, classwise and canonical calibration error are used to

evaluate per-structure and overall calibration levels. Derived from individual channel masks, the classwise CE in multi-class segmentation simplifies to binary CE for each channel. In addition, Popordanoska et al. (2021) proves that the absolute value of volume bias (V-Bias) is upper-bounded by CE. Many calibration methods like temperature scaling (Guo et al., 2017) and isotonic regression (Zadrozny & Elkan, 2002) have been proposed to improve the calibration of classification scores. However, no previous work analyzes how miscalibration affects downstream ratio-based estimates.

3 METHODS

We begin with relevant definitions in Sec. 3.1 to establish the theoretical background. In Sec. 3.2, we present our main contribution, the uncertainty decomposition and corresponding confidence intervals.

3.1 PRELIMINARIES

The ratio-based biomarker is clinically defined as the ratio between two volumes V_A and V_B (Henker et al., 2019; 2017). We consider the ratio estimation within a standard segmentation framework, where V_A and V_B are calculated from predicted probabilities.

Definition 3.1 (Ratio from Segmentation Networks). Given per-pixel inputs $\{z_i\}_{i=1}^n$, labels $\{y_{A,i}, y_{B,i}\}_{i=1}^n$ and segmentation model $g: z_i \rightarrow g_A(z_i), g_B(z_i) \in [0, 1]$, the labeled ratio r_{gt} and predicted ratio $\hat{r}$ within n pixels are calculated by:

$$r_{\text{gt}} = \frac{\bar{y}_A}{\bar{y}_B} = \frac{\sum_{i=1}^n y_{A,i}}{\sum_{i=1}^n y_{B,i}}, \text{ and } \hat{r} = \frac{\bar{g}_A}{\bar{g}_B} = \frac{\sum_{i=1}^n g_A(z_i)}{\sum_{i=1}^n g_B(z_i)}. \quad (1)$$

Proposition 3.2 (Conformal Prediction for Regression (Shafer & Vovk, 2008)). *Given groundtruth r_{gt} , prediction $\hat{r}$ and the absolute error residual $\text{AE}_r := |r_{\text{gt}} - \hat{r}|$, let $q_{r,\delta}$ denote the $1 - \delta$ quantile of the instance-wise AE_r on a validation (calibration) set. Then, with probability at least $1 - \delta$*

$$r_{\text{gt}} \in [\hat{r} - q_{r,\delta}, \hat{r} + q_{r,\delta}], \quad (2)$$

From Def. 3.1, the predicted ratio $\hat{r}$ is determined by the probability volumes predicted by the network. Since the network is not perfectly calibrated, quantifying the uncertainty in its predictions is closely tied to assessing the uncertainty of the derived biomarker. To this end, we introduce two metrics: volume bias and the calibration error.

Definition 3.3 (Volume Bias (Popordanoska et al., 2021)). Given a segmentation model $g: \mathcal{Z} \rightarrow [0, 1]$ that predicts the probability of $y \in \{0, 1\}$, the volume bias is defined as:

$$\text{V-Bias}(g) := \mathbb{E}_{(z,y) \sim P} [g(z) - y]. \quad (3)$$

Definition 3.4 (Calibration Error (Kumar et al., 2019)). Given a model $g: \mathcal{Z} \rightarrow [0, 1]$ that predicts the probability of $y \in \{0, 1\}$, the calibration error is defined as:

$$\text{CE}(g) := \mathbb{E}_{(z,y) \sim P} [|g(z) - \mathbb{E}[y = 1 | g(z)]|], \quad (4)$$

Proposition 3.5 (The Relationship of V-Bias and CE (Popordanoska et al., 2021)). *Given segmentation model $g: \mathcal{Z} \rightarrow [0, 1]$, the absolute value of volume bias is upper bound by the calibration error, i.e., $|\text{V-Bias}(g)| \leq \text{CE}(g)$.*

Leveraging the upper bound relationship and statistical properties, we decompose the uncertainty of the ratio estimation pipeline and give respective confidence intervals in Sec. 3.2.

3.2 CARE: CONFIDENCE-AWARE RATIO ESTIMATION

In this section, we illustrate our insight of uncertainty analysis based on two key observations, as shown in Fig. 3. The first observation is that the ratio estimator $\hat{r} = \frac{\bar{y}}{\bar{x}}$ is subject to instance-wise randomness, which we capture using statistical tools such as Markov’s inequality to derive an *estimation-based interval*. The second observation is that the network is not perfectly calibrated, introducing a global, model-level error affecting both the numerator and denominator; this gives rise to the *calibration-based interval*. Combining these two sources yields the overall uncertainty bound.

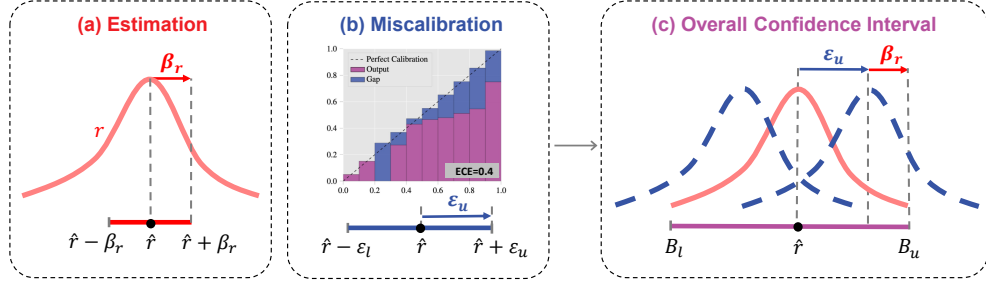


Figure 3: **Our confidence interval considering estimation and miscalibration.** (a) shows Markov bounds from the estimator. (b) illustrates the prediction offset $\epsilon_{l,u}$ due to miscalibration. (c) is the overall confidence interval $r \in [B_l, B_u]$.

Estimation-based interval. Van Kempen & Van Vliet (2000) provides an approximated theoretical result for ratio statistics. However, their derivation critically relies on the assumption that the addends in $\bar{x}$ and $\bar{y}$ are independent. Therefore, the result in Van Kempen & Van Vliet (2000) is not directly applicable in imaging analysis for violating spatial patterns. As a remedy, we construct Markov bounds as an estimation-based confidence interval for $\hat{r}$ using Markov inequality (Resnick, 2003). Although this approach leads to more conservative bounds, it avoids strong assumptions such as pixel independence, making it more applicable to image data.

Proposition 3.6 (Estimation-based Confidence Interval). *Given an estimator $\hat{r} = \frac{\bar{y}}{\bar{x}}$ of the fraction $r = \frac{\mathbb{E}[y]}{\mathbb{E}[x]}$ with random variables x and y , it holds with at least $1 - \alpha$ probability that*

$$r \in [\hat{r} - \beta_{r,\alpha}, \hat{r} + \beta_{r,\alpha}], \quad (5)$$

where $\beta_{r,\alpha} := \frac{\sqrt{\text{SE}_{\hat{r}}}}{\sqrt{\alpha}}$ is the half-width of the bound, and $\text{SE}_{\hat{r}} := \mathbb{E}[(\hat{r} - r)^2]$ is the expected squared error.

Then we conduct a Taylor expansion of $\text{SE}_{\hat{r}}$ to receive an approximation we can estimate in practice.

Proposition 3.7. *Assume all central moments of the independently and identically distributed random variables $(x_1, y_1), \dots, (x_n, y_n) \sim \mathbb{P}_{xy}$ in the estimator $\hat{r} = \frac{\bar{y}}{\bar{x}}$ exist, then we have*

$$\text{SE}_{\hat{r}} = \frac{1}{n} \left(\frac{\text{Var}(y)}{\mu_x} + \text{Var}(x) \frac{\mu_y^2}{\mu_x^4} - 2 \text{Cov}(x, y) \frac{\mu_y}{\mu_x^3} \right) + O\left(\frac{1}{n^2}\right). \quad (6)$$

The proof is given in the appendix. Then the estimator is:

$$\widehat{\text{SE}}_{\hat{r}} := \frac{1}{n} \left(\frac{\hat{\sigma}_y^2}{\bar{x}} + \frac{\hat{\sigma}_x^2 \bar{y}^2}{\bar{x}^4} - 2 \frac{\hat{\sigma}_{xy} \bar{y}}{\bar{x}^3} \right), \quad (7)$$

with the sample variances $\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$, $\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$, and sample covariance $\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$. Under i.i.d. assumption, the estimator $\widehat{\text{SE}}_{\hat{r}}$ is consistent, i.e., $\widehat{\text{SE}}_{\hat{r}} \rightarrow \text{SE}_{\hat{r}}$ in probability for $n \rightarrow \infty$. The proof is presented in the appendix B.1.

Calibration-based interval. The estimation-based bounds involve local uncertainty that stems from statistical properties. Then we analyze the second source of uncertainty: volume bias caused by miscalibration. Inspired by Prop. 3.2, we propose a fine-grained calibration-based confidence interval by considering the uncertainty of target (A) and RoI (B) regions separately. Unlike vanilla conformal prediction, where analysis starts from the final $\hat{r}$, we adopt quantiles of V_A and V_B and report the corresponding interval as CARE (V-Bias). As described in Proposition 3.5, V-Bias of target (A) and RoI (B) regions is upper bounded by their calibration errors, i.e., $|\text{V-Bias}(g_A)| \leq \text{CE}(g_A)$, $|\text{V-Bias}(g_B)| \leq \text{CE}(g_B)$. This motivates the more conservative interval named as CARE (ECE).

Proposition 3.8 (Calibration-based Confidence Interval). *Consider a segmentation model $g(z) = (g_A(z), g_B(z))$ with the random variable z representing pixel inputs of instance I , and targets y_A*

and y_B . On a validation (calibration) set $\mathcal{D}_{cal}$, define $q_{A,\delta/2}$ and $q_{B,\delta/2}$ as the $1 - \delta/2$ quantile of the instance-wise volume bias or calibration errors of g_A and g_B . Then, it holds with at least $1 - \delta$ probability that

$$\frac{\mathbb{E}[y_A | I]}{\mathbb{E}[y_B | I]} \in \left[\frac{\mathbb{E}[g_A(z) | I]}{\mathbb{E}[g_B(z) | I]} - \epsilon_{l,\delta}, \frac{\mathbb{E}[g_A(z) | I]}{\mathbb{E}[g_B(z) | I]} + \epsilon_{u,\delta} \right], \quad (8)$$

where $\epsilon_{l,\delta} := \frac{\mathbb{E}[g_A(z)]}{\mathbb{E}[g_B(z)]} - \frac{\mathbb{E}[g_A(z)] - q_{A,\delta/2}}{\mathbb{E}[g_B(z)] + q_{B,\delta/2}}$, $\epsilon_{u,\delta} := \frac{\mathbb{E}[g_A(z)] + q_{A,\delta/2}}{\mathbb{E}[g_B(z)] - q_{B,\delta/2}} - \frac{\mathbb{E}[g_A(z)]}{\mathbb{E}[g_B(z)]}$ are the widths of the lower and upper calibration bounds, respectively.

The proof is presented in the appendix B.2. In experiments, CARE (V-Bias) takes the quantile of IV-Bias as $q_{N,T}$ while CARE (ECE) considers ECE (Guo et al., 2017) quantiles. To combine both intervals, we make the following statement, which is analogous to multiple testing. This way, we can consider both uncertainties in practice.

Proposition 3.9 (Overall Confidence Interval). Assume we have a ratio estimator $\hat{r} = \frac{\sum_i g_A(z_{i,I})}{\sum_i g_B(z_{i,I})}$ for pixel measurements $\{z_{i,I}\}_{i=1}^n$ of an instance I based on neural network outputs $g(z_{i,I}) = (g_A(z_{i,I}), g_B(z_{i,I}))$. Let y_A and y_B be the instance-wise target random variables. Then, it holds with at least $1 - \alpha - \delta$ probability that

$$\frac{\mathbb{E}[y_A | I]}{\mathbb{E}[y_B | I]} \in \left[\frac{\sum_i g_A(z_{i,I})}{\sum_i g_B(z_{i,I})} - \epsilon_{l,\delta} - \beta_{r,\alpha}, \frac{\sum_i g_A(z_{i,I})}{\sum_i g_B(z_{i,I})} + \epsilon_{u,\delta} + \beta_{r,\alpha} \right], \quad (9)$$

where $\beta_{r,\alpha}$ is defined as in Prop. 3.6 and $\epsilon_{l,\delta}, \epsilon_{u,\delta}$ as in Prop. 3.8.

The interval width $w = B_u - B_l$ measures the uncertainty level, as a result, a wide interval over thresholds alarms for manual examination. In the experiments, we alternate through various α and δ for a fixed $\alpha + \delta$ with grid search to observe the impact on the interval width. This way, we can choose the smallest interval under a desired coverage rate.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on two brain tumor segmentation datasets: MSD-Task01 (Antonelli et al., 2022) and BraTS21 (Baid et al., 2021), which include 484 and 1251 MRI volumes, respectively, with four modalities (T1, T2, T1ce, FLAIR) and four annotations (edema, necrosis, enhancing tumor, background). Their necrosis-to-tumor ratio (NTR) is defined as the ratio of necrosis V_N to the whole tumor area V_T (edema, necrosis, enhancing). In addition, we include KiTS23 (Myronenko et al., 2023), a CT dataset of 489 kidney volumes. Its tumor-to-kidney ratio (TKR) is defined as $\frac{V_{tumor}}{V_{whole\ kidney}}$. A nested five-fold cross-validation is used for all datasets. In the outer loop, four folds are used for training and validation, and the remaining one fold for testing. Within the inner loop, 10% of the training data is held out as a validation set $\mathcal{D}_{cal}$ to estimate the quantile of V-Bias and ECE. Predicted ratio $\hat{r}$ and labeled ratio r_{gt} are calculated from Def. 3.1.

Segmentation models. We conduct experiments using nnUNet (Isensee et al., 2021), nnFormer (Zhou et al., 2021) and UNETR++ (Zhou et al., 2021). All models are trained using four-modality MRI scans, label-based supervision and softmax activation under a single A100 GPU. We investigate different loss functions as ensemble baselines: cross-entropy (XE) (Bishop & Nasrabadi, 2006), soft Dice (SD) (Milletari et al., 2016), TopK (Deng et al., 2009), and their combinations (XE-SD, Top10-SD), while use XE-SD for the main results.

UQ baselines. To control the confidence level to be $C = 0.68$, we adopt a quantile for each baseline method. For conformal prediction (Vovk et al., 1999; Papadopoulos et al., 2002) and CARE, we obtain quantiles from the validation set. Specifically, for conformal prediction we take the 0.68 quantile (0.68Q) of AE_r from the validation set as the half-width (Prop. 3.2), while for CARE we adopt dynamic ECE quantiles or V-Bias quantiles by conducting a grid search under the constraint of $1 - \alpha - \beta = 0.68$ (Prop. 3.9). To implement resampling, we repeatedly sample pixels from an instance and calculate its ratio estimate for 100 times, then adopt the $[0.16Q, 0.84Q]$ from 100 repetitions as the 0.68 confidence level. For a volume of N pixels, we take $0.1N$ random pixels each

Table 1: **Comparison of the coverage guarantee on MSD-Task01 dataset** ($C = 0.68$). We report the overall coverage rate (%) on test-set. CARE always satisfies the desired confidence level, while other methods fall below in most cases.

Coverage (%)	nnUNet _{2d}	nnUNet _{3d}	nnFormer	UNETR++
Subsampling	6.19 $\pm$ 0.77	9.28 $\pm$ 0.92	5.74 $\pm$ 0.72	8.22 $\pm$ 0.91
Bootstrap	5.34 $\pm$ 0.61	8.18 $\pm$ 0.62	5.53 $\pm$ 0.75	8.12 $\pm$ 0.71
Conformal prediction	71.34 $\pm$ 2.00	67.01 $\pm$ 3.57	67.39 $\pm$ 1.66	65.75 $\pm$ 2.16
CARE (V-Bias)	93.61 $\pm$ 1.14	86.60 $\pm$ 1.49	81.92 $\pm$ 1.31	76.43 $\pm$ 2.21
CARE (ECE)	94.22 $\pm$ 0.99	93.61 $\pm$ 0.71	87.94 $\pm$ 0.97	89.58 $\pm$ 1.02

time without replacement for subsampling (Politis & Romano, 1994), and sample N pixels with replacement each time for bootstrapping (Mooney et al., 1993). For Bayesian methods, conducting numerous forward passes to estimate a “tunable” quantile is computationally impractical; thus, we report the results of three standard deviations (3σ).

Metrics. We evaluate the performance of various methods across four criteria: i) *Coverage guarantee*: ability to achieve the desired confidence level, quantified by coverage rate, ii) *Adaptiveness*: capacity to capture sample variability (e.g., prediction error) and segmentation difficulty (e.g., tumor size); iii) *Tunability*: flexibility to choose a user-specified confidence level; iv) *Practical deployment*: whether the method operates as a plug-in module (non-intrusive to the model architecture) and maintains computational efficiency (without requiring multiple sampling or repeated inference steps).

4.2 RESULTS

We demonstrate the claimed properties in Sec. 1 of our method: coverage guarantee, adaptiveness, tunability and practical deployment characteristics (plug-in compatibility and computational efficiency). Moreover, we analyze the uncertainty source to get an insight into the dominant component.

Coverage guarantee. As described in Sec.1, a conservative confidence interval achieves coverage probability higher than the nominal confidence level, i.e., achieving over 68% coverage when aiming for 68% confidence level. We report coverage rate (%) of different UQ methods at 0.68 confidence level in Table 1, which measures *the proportion of samples whose true values fall within the confidence intervals*. Empirically, our intervals show higher likelihoods of satisfying the prescribed confidence level of 0.68 compared with other UQ methods. Considering the suboptimal performance of sampling-based methods, our following comparison focuses on CP and CARE.

Adaptiveness. Beyond achieving the guaranteed coverage rate, the confidence interval should be sample-adaptive to identify unreliable predictions effectively. We demonstrate this capability by examining the “dataset-level interval” distribution of MSD-Task01 in Fig. 4. As observed, CP values lie within a narrow range and thus fail to effectively indicate which samples are unreliable. In contrast, our method produces intervals that vary significantly in width. Given an interval width threshold, our method can effectively trigger alarms for cases with wide intervals (indicating high uncertainty), thereby overcoming CP’s limitation of producing uniformly narrow confidence ranges.

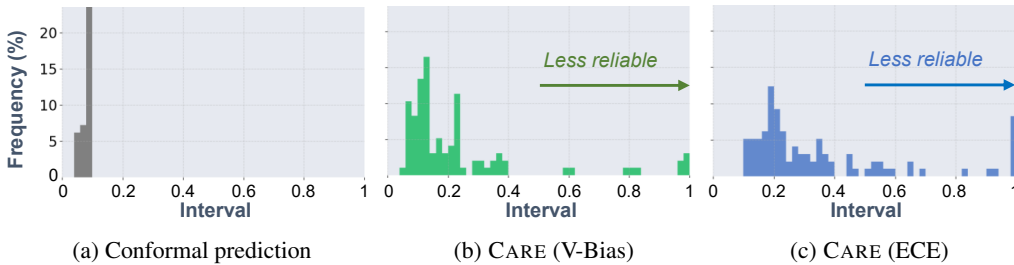


Figure 4: **Comparison of interval distribution on MSD-Task01 dataset** ($C = 0.68$). We report the frequency histogram of NTR intervals in test-set, where CARE triggers a human-check alarm when the interval is too wide.

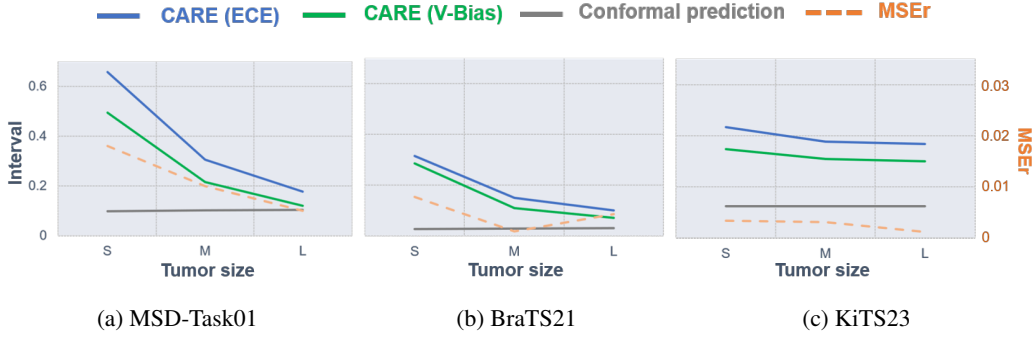


Figure 5: **Comparison of tumor-size adaptiveness on nnUNet_{3d}** ($C = 0.68$). For each dataset, we report the average interval width in three groups categorized by tumor sizes. Intuitively, interval width should reflect the MSE_r tendency in $---$. Compared with the indistinguishable results of CP, CARE becomes wider for small tumors (hard samples) and tighter for large ones (simple samples).

Furthermore, the uncertainty should correlate appropriately with segmentation difficulty. For instance, small tumors are hard to detect and segment for their small size, low contrast and susceptibility to noise. Empirically, hard samples with small sizes or blurry boundaries tend to yield erroneous predictions (large mean squared error), necessitating wider intervals to ensure coverage. To validate this adaptive behavior, we present fine-grained analysis of MSE_r (error measures) and interval width (uncertainty measures) in Fig. 5. Specifically, we report NTR for Fig. 5a and 5b, and report TKR for Fig. 5c. We stratify tumors into small (S), medium (M), and large (L) categories based on the $\frac{1}{3}$ and $\frac{2}{3}$ quantiles of tumor sizes in test-set. As illustrated, our intervals widths are proportional to the segmentation difficulty: smaller, more challenging tumors receive wider intervals, while larger, easier-to-segment tumors receive narrower intervals.

Tunability. CARE offers two variants that allow clinicians to select either conservative or informative bounds by choosing CARE (ECE) or (V-Bias). To demonstrate tunability and coverage guarantee across different confidence levels, we report coverage rates for varying confidence thresholds on two biomarkers: NTR and CTR in Fig. 6. The coverage rate is expected to increase proportionally with the increased confidence level. However, conformal prediction shows significant limitations: it only achieves adequate coverage at isolated points ($C = 0.7$ for NTR and $C = 0.6, 0.7$ for CTR), while falling below the target confidence level at all other tested thresholds. Additionally, conformal prediction consistently fails to provide adequate coverage for small tumors (NTR-S) across the entire confidence range, as demonstrated in Fig. 6c. In contrast, both variants of our method consistently achieve coverage rates above the desired confidence level.

Other baselines. Bayesian UQ methods like deep ensemble (Lakshminarayanan et al., 2017) and Monte Carlo dropout (Srivastava et al., 2014) require modifications to model architectures or training procedures, in contrast to the previously discussed plug-in methods. For practical usage, CP and

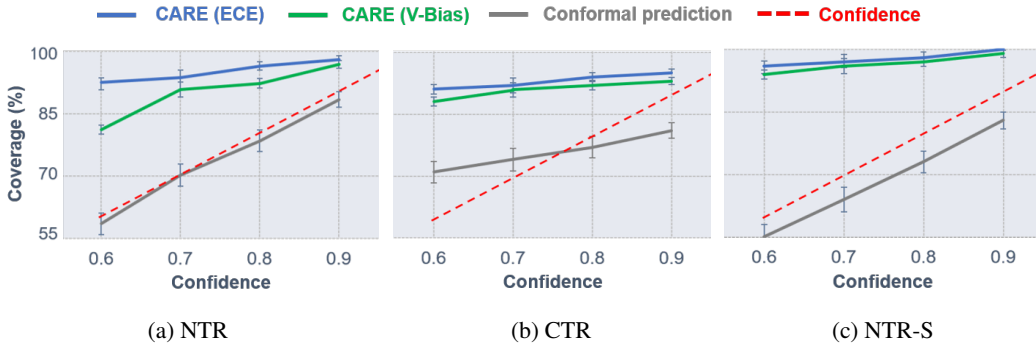


Figure 6: **Performance comparison across confidence levels on MSD-Task01 and nnUNet_{3d}**. Methods over $---$ achieve the desired coverage. CARE is tunable to maintain the target coverage at any confidence level and outperforms CP in detecting high-risk small tumors (NTR-S).

CARE achieve linear computational complexity with minimal computational overhead. In comparison, Bayesian methods face significant practical limitations in clinical settings. Performing numerous ensemble predictions or dropout inferences is computationally expensive and often impractical for real-time applications.

In Table 2, we implement two ensemble configurations on BraTS21 using nnUNet_{3d}: i) training with five random seeds, and ii) training with five different loss functions (XE, SD, Top10, XE-SD, Top10-SD). For dropout configurations, we add dropout layers to the nnUNet decoder with dropout probabilities of 0.3 and 0.5, and perform 20 stochastic forward passes. The interval widths of Bayesian methods are set as 3σ to provide a very conservative interval. Nevertheless, we find that the intervals of Bayesian methods are still too narrow to provide valid intervals, most likely due to the lack of an appropriate prior (Noci et al., 2021).

Uncertainty source analysis. As described in Sec. 3.2, we decompose uncertainty into miscalibration and intrinsic bias of ratio estimation. We validate this empirically by analyzing 10 randomly selected volumes from BraTS21, calculating ECE-based (I_{ECE}) and overall CARE confidence intervals (I). The results in Fig. 7 show that I_{ECE} dominates the overall interval I , indicating that model miscalibration is the primary uncertainty source in ratio estimation.

Discussion about post-hoc calibration We calculate ECE and our intervals on different temperatures in appendix A.2. For overconfident models, temperature_{>1} helps for better calibrated performance. When ECE decreases, our interval becomes tighter consistently.

5 CONCLUSION

We propose CARE, a confidence-aware framework for estimating ratio-based biomarkers from segmentation network outputs. Our method addresses a common limitation of prior works that focus solely on point estimates without confidence guarantees. We disentangle two key sources of uncertainty, *i.e.* network prediction error and statistical bias. Our empirical findings highlight that miscalibration is a dominant contributor to uncertainty. Our framework offers several practical advantages: it operates as a model-agnostic plugin module, provides sample-level adaptive uncertainty estimates in a single forward pass without requiring multiple sampling, and allows users to flexibly adjust confidence levels. In summary, this work represents an important step toward trustworthy deployment of deep learning in clinical settings by providing practitioners with both accurate biomarker estimates and reliable confidence bounds.

Limitations and future work. Despite the practical advantages, our work has several limitations. First, we assume that the validation and test sets are drawn from the same distribution. Although it is standard in supervised learning settings, but may not hold under domain shifts. In practice, domain shifts arise due to differences in scanners, acquisition protocols, or patient populations. As a result, our confidence interval may not remain valid in these scenarios. Addressing this challenge with label-free calibration error estimators (e.g. Wang et al. (2020); Popordanoska et al. (2024)) is a promising direction for future work. Second, the calibration quality of the underlying segmentation network has an impact on the tightness of the derived confidence intervals. Specifically, when the calibration error is large, the resulting confidence intervals may become overly conservative. Improving calibration in segmentation networks would directly translate into narrower, more informative confidence intervals within our approach. Finally, while our framework shows good performance on public datasets, clinical validation is needed to assess its real-world impact on decision-making and patient outcomes.

Table 2: **Comparison of UQ methods on BraTS21 with nnUNet_{3d}.** Ensemble and dropout methods provide too narrow NTR intervals.

	Interval	Coverage (%)
Ensemble _{loss}	0.088 ± 0.003	46.03 ± 1.21
Ensemble _{seed}	0.041 ± 0.002	43.03 ± 1.13
Dropout _{0.3}	0.033 ± 0.001	27.09 ± 1.02
Dropout _{0.5}	0.038 ± 0.001	29.63 ± 1.03

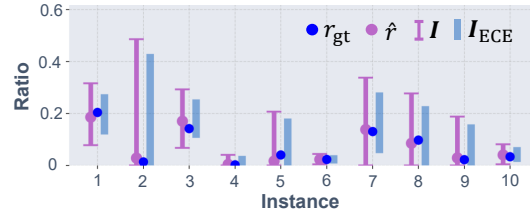


Figure 7: **Uncertainty decomposition of 10 samples.** Miscalibration is the main contributor to the overall uncertainty.

REFERENCES

- Akmalbek Bobomirzaevich Abdusalomov, Mukhridin Mukhiddinov, and Taeg Keun Whangbo. Brain tumor detection based on deep learning approaches and magnetic resonance imaging. *Cancers*, 15(16):4172, 2023.
- Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.
- Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), 2022.
- Majd Bahna, Muriel Heimann, Christian Bode, Valeri Borger, Lars Eichhorn, Erdem Güresir, Motaz Hamed, Ulrich Herrlinger, Yon-Dschun Ko, Felix Lehmann, et al. Tumor-associated epilepsy in patients with brain metastases: necrosis-to-tumor ratio forecasts postoperative seizure freedom. *Neurosurgical Review*, 45(1):545–551, 2022.
- Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.
- Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- David A Freedman. Bootstrapping regression models. *The annals of statistics*, pp. 1218–1228, 1981.
- Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pp. 1321–1330. PMLR, 2017.
- Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284, 2021.
- Christian Henker, Thomas Kriesen, Anne Glass, Björn Schneider, and Jürgen Piek. Volumetric quantification of glioblastoma: experiences with different measurement techniques and impact on survival. *Journal of neuro-oncology*, 135:391–402, 2017.
- Christian Henker, Marie Cristin Hiepel, Thomas Kriesen, Moritz Scherer, Anne Glass, Christel Herold-Mende, Martin Bendszus, Sönke Langner, Marc-André Weber, Björn Schneider, et al. Volumetric assessment of glioblastoma and its predictive value for survival. *Acta Neurochirurgica*, 161:1723–1732, 2019.
- Brian R Herts, Deirdre M Coll, Andrew C Novick, Nancy Obuchowski, Grant Linnell, Susan L Wirth, and Mark E Baker. Enhancement characteristics of papillary renal neoplasms revealed on triphasic helical ct of the kidneys. *American Journal of Roentgenology*, 178(2):367–372, 2002.

- David Joon Ho, Narasimhan P Agaram, Peter J Schüffler, Chad M Vanderbilt, Marc-Henri Jean, Meera R Hameed, and Thomas J Fuchs. Deep interactive learning: an efficient labeling approach for deep learning-based osteosarcoma treatment response assessment. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V* 23, pp. 540–549. Springer, 2020.
- Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- Hamed Karimi and Reza Samavi. Quantifying deep learning model uncertainty in conformal prediction. In *Proceedings of the AAAI Symposium Series*, volume 1, pp. 142–148, 2023.
- Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.
- Min Suk Kim, Soo-Yong Lee, Wan Hyeong Cho, Won Seok Song, Jae-Soo Koh, Jun Ah Lee, Ji Young Yoo, and Dae-Geun Jeon. Tumor necrosis rate adjusted by tumor volume change is a better predictor of survival of localized osteosarcoma patients. *Annals of surgical oncology*, 15: 906–914, 2008.
- Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7–11, 2015, Proceedings, Part I* 15, pp. 68–85. Springer, 2015.
- Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *NeurIPS*, volume 32, 2019.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.
- Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.
- Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage, 1993.
- Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pp. 1–7. 2023.
- Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep relu networks. In *NeurIPS*, volume 34, 2021.
- OpenAI. Chatgpt (gpt-5). <https://chat.openai.com>, 2025. Large language model used for text polishing and clarification.
- Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings* 13, pp. 345–356. Springer, 2002.
- Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pp. 2031–2050, 1994.

- Teodora Popordanoska, Jeroen Bertels, Dirk Vandermeulen, Frederik Maes, and Matthew B Blaschko. On the relationship between calibrated predictors and unbiased volume estimation. In *MICCAI*, pp. 678–688, 2021.
- Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *NeurIPS*, 35:7933–7946, 2022.
- Teodora Popordanoska, Gorjan Radevski, Tinne Tuytelaars, and Matthew Blaschko. Lascal: Label-shift calibration without target labels. *Proceedings NeurIPS 2024*, 2024.
- Sidney Resnick. *A probability path*. Springer Science & Business Media, 2003.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.
- Axel-Jan Rousseau, Thijs Becker, Simon Appeltans, Matthew Blaschko, and Dirk Valkenburg. Post hoc calibration of medical segmentation models. *Discover Applied Sciences*, 7(3):180, 2025.
- Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.
- Michael Spivak. *Calculus*. Cambridge University Press, 2006.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467. PMLR, 2019.
- GMP Van Kempen and LJ Van Vliet. Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry: The Journal of the International Society for Analytical Cytology*, 39(4): 300–305, 2000.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.
- Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.
- Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23652–23662, 2023.
- Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *European Conference on Computer Vision*, pp. 456–472. Springer, 2022.
- Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020.
- Huifen Ye, Yunrui Ye, Yiting Wang, Tong Tong, Su Yao, Yao Xu, Qingru Hu, Yulin Liu, Changhong Liang, Guangyi Wang, et al. Automated assessment of necrosis tumor ratio in colorectal cancer using an artificial intelligence-based digital pathology analysis. *Medicine Advances*, 1(1):30–43, 2023.
- Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.
- Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pp. 11117–11128. PMLR, 2020.

Hancheng Y. Zhou, Jian Guo, Y. Zhang, et al. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

Appendix

In Appendix A, we present additional experimental results, relevant to our methodology and in support of the main paper. In Appendix B, we offer the proofs of propositions in the main paper. Finally, we claim the LLM usage in Appendix C.

A MORE EMPIRICAL RESULTS

This section presents additional empirical results. To further examine different components of our framework, we first provide more visualization in Sec. A.1 to further demonstrate the coverage guarantee and adaptiveness of our confidence intervals. In the main paper, we observe that miscalibration is the main cause of uncertainty. As an extension, we conduct post-hoc calibration in Sec. A.2 to report ECE and our confidence interval under different temperatures. Finally, in Sec. A.3, we replace the default ECE metric (Guo et al., 2017) with KDE-based calibration error (Popordanoska et al., 2022), highlighting the flexibility of our framework with respect to calibration error estimators. All experiments are conducted at the 0.68 confidence level.

A.1 COVERAGE GUARANTEE AND ADAPTIVENESS

In the main paper, we just visualize the confidence intervals of 10 randomly selected samples in Fig. 7. To provide a more comprehensive, “bird-eye” view of our method’s behavior, we extend this analysis to the whole test samples in Fig. A, where we plot r_{gt} and the confidence intervals I under three methods. For clarity, the sample indices are omitted. As shown here, the ground-truth ratio r_{gt} (blue point) always lies well within our predicted confidence interval, while for conformal prediction, r_{gt} occasionally falls outside the interval when the upper or lower bounds are too narrow. The conservative property of CARE is particularly important in clinical settings to provide a reliable and informative reference.

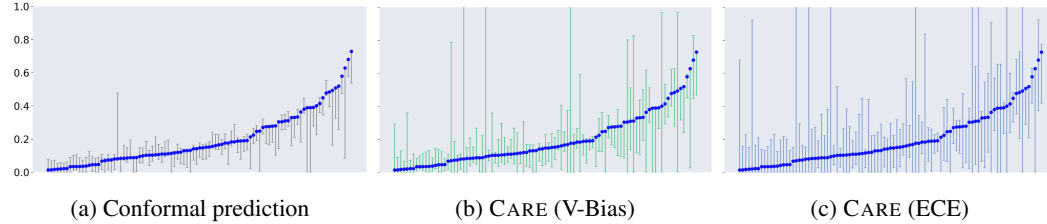


Figure A: **Visualization of our confidence intervals on MSD and nnUNet_{3d}.** The x-axis represents all test samples sorted by labeled ratio r_{gt} , and the y-axis displays the valid range of ratio estimates. As an extension of Fig. 5, we show our adaptiveness towards tumor sizes by three different pretrained models in Fig. B. As the previous setting, all samples are divided into three groups: small (S), medium (M) and large (L). Since CARE is a plug-in module with the model-agnostic nature, the adaptiveness holds on all pretrained models.

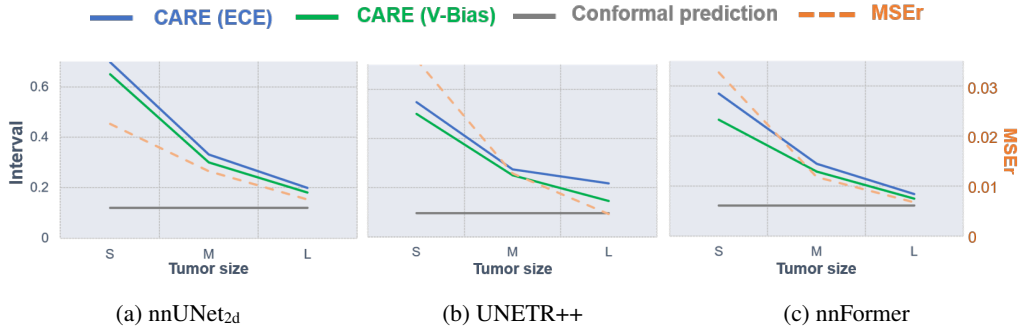


Figure B: **Comparison of tumor-size adaptiveness on MSD.** CARE is adaptive to different tumor sizes across all models.

A.2 POST-HOC CALIBRATION

Our main contribution is the confidence interval for ratio estimation. Nevertheless, we acknowledge that the calibration property is very important for downstream tasks, as demonstrated by Fig. 7. In this section, we report different temperature parameters on a pretrained nnUNet_{3d}, to observe the effect of post-hoc calibration. We report the ECE of necrosis and tumor ($ECE_{N,T}$) and average interval width of CARE (ECE) under different temperatures in Table A. For illustration, we scale up ECE by 100. As observed, both ECE and our interval width decrease as the temperature increases. The consistency demonstrates that better-calibrated models have tighter confidence intervals.

Table A: **Comparison of different temperature parameters on MSD and nnUNet_{3d}**. When the temperature moves towards better calibration ($ECE \downarrow$), our interval becomes narrower (Interval $\downarrow$).

Temperature	ECE_N	ECE_T	CARE (ECE)
0.1	0.112 ± 0.004	0.166 ± 0.006	0.405 ± 0.023
0.5	0.101 ± 0.005	0.141 ± 0.004	0.364 ± 0.012
1.0	0.098 ± 0.006	0.134 ± 0.003	0.355 ± 0.025
1.5	0.097 ± 0.004	0.131 ± 0.006	0.353 ± 0.027
2.0	0.097 ± 0.002	0.129 ± 0.004	0.353 ± 0.023
3.0	0.096 ± 0.029	0.128 ± 0.005	0.352 ± 0.026
4.0	0.096 ± 0.003	0.127 ± 0.007	0.351 ± 0.024
8.0	0.095 ± 0.008	0.126 ± 0.004	0.349 ± 0.035

A.3 OTHER CALIBRATION ERROR ESTIMATORS

We replace the calibration error estimator ECE (Guo et al., 2017) with ECE_{kde} (Popordanoska et al., 2022) for comparison. In segmentation task, we use binary calibration error which corresponds to Beta kernel in ECE_{kde} . Since kernel computation is much expensive than bins, and all pixels together will cause OOM error, we sample 10^4 pixels once for ECE_{kde} estimation, and repeat 5 times to report the average value. Observed from 10 volumes in Fig. C, ECE_{kde} tends to provide wider bounds, which is suitable for conservative preference. Notably, the estimator is flexible to plug into our framework. We don't aim to give any recommendations, depending on the priority of tightness or informativeness. For a conservative estimator, the alarm thresholds of the interval width should also increase to avoid over-checking. As future work, these differentiable ECE estimator may facilitate obtaining tighter confidence intervals through carefully designed optimization.

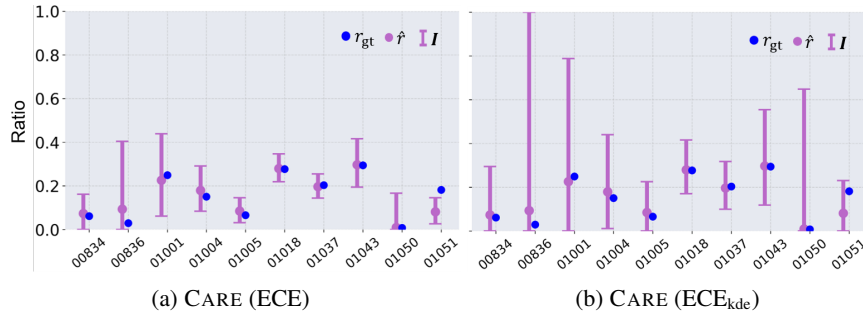


Figure C: **Comparison of different calibration error estimators on BraTS21 and nnUNet_{3d}**. Adopting (a) ECE (Guo et al., 2017) is generally tighter and more informative than (b) KDE.

B PROOFS

In this section, we give the corresponding proof of Markov bounds (B.1) and miscalibration bounds (B.2) mentioned in Sec. 3.2 of the main paper. In addition, we derive a debiased estimator in Sec. B.3.

B.1 MARKOV BOUNDS

Van Kempen & Van Vliet (2000) provides a confidence interval of the ratio estimator $\frac{\bar{y}}{\bar{x}}$ based on asymptotic normal assumptions and by using the variance $\sigma_r^2 := \text{Var}(\frac{\bar{y}}{\bar{x}})$. However, adopting their results assumes that all pixels are independently and identically distributed, i.e., $(x_1, y_1), \dots, (x_n, y_n) \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{xy}$. In addition, they perform multiple approximation steps, and some approximations happen within the square operator. How the estimator behaves facing a violation of these assumptions is unknown in practice. In the following, we prove the alternative approach, we proposed in the main paper, which is based on Markov's inequality (Resnick, 2003). For conciseness, the “ $\approx$ ” sign is avoided while we directly note the remainder terms for a rigorous analysis.

To avoid relying on any distribution assumptions, we construct a confidence interval via Markov's inequality for the estimator $\hat{r} = \frac{\bar{y}}{\bar{x}}$ and target $r = \frac{\mu_y}{\mu_x}$. We have

$$\mathbb{P}(|\hat{r} - r| \geq k\sqrt{\text{SE}_{\hat{r}}}) = \mathbb{P}((\hat{r} - r)^2 \geq k^2 \text{SE}_{\hat{r}}) \leq \frac{1}{k^2} \quad (10)$$

with the squared error $\text{SE}_{\hat{r}} := \mathbb{E}[(\hat{r} - r)^2]$. We emphasize that in general $\sqrt{\text{SE}_{\hat{r}}} \neq \sigma_r$.

In main paper, we denote $\alpha := \frac{1}{k^2}$ as the non-coverage probability. For instance, adopting the $1 - \alpha = 75\%$ confidence interval corresponds to $\alpha = \frac{1}{k^2} = 0.25$ or $k = 2$. Then the half-width of confidence interval is $2\sqrt{\text{SE}_{\hat{r}}}$, i.e., two times the root squared error. This is more conservative than using the normal assumption, but requires no distribution assumption.

Now, we compute the squared error via Taylor expansion (Spivak, 2006). First, note that

$$\text{SE}_{\hat{r}} = \mathbb{E}\left[\left(\frac{\bar{y}}{\bar{x}} - \frac{\mu_y}{\mu_x}\right)^2\right] = \mathbb{E}\left[\frac{\bar{y}^2}{\bar{x}^2}\right] - 2\frac{\mu_y}{\mu_x}\mathbb{E}\left[\frac{\bar{y}}{\bar{x}}\right] + \frac{\mu_y^2}{\mu_x^2}. \quad (11)$$

We perform a Taylor expansion of $\frac{\bar{y}^2}{\bar{x}^2}$ around $\frac{\mu_y}{\mu_x}$ to compute its expectation:

$$\begin{aligned} \frac{\bar{y}^2}{\bar{x}^2} &= \frac{\mu_y^2}{\mu_x^2} + 2(\bar{y} - \mu_y)\frac{\mu_y}{\mu_x^2} - 2(\bar{x} - \mu_x)\frac{\mu_y^2}{\mu_x^3} \\ &\quad + (\bar{y} - \mu_y)^2\frac{1}{\mu_y} + 3(\bar{x} - \mu_x)^2\frac{\mu_y^2}{\mu_x^4} - 4(\bar{y} - \mu_y)(\bar{x} - \mu_x)\frac{\mu_y}{\mu_x^3} \\ &\quad + \sum_{i,j: i+j \geq 3} (\bar{y} - \mu_y)^i (\bar{x} - \mu_x)^j \frac{\partial^{i+j}}{\partial^i \mu_y \partial^j \mu_x} \frac{\mu_y^2}{\mu_x^2} \end{aligned} \quad (12)$$

from which follows

$$\begin{aligned} \mathbb{E}\left[\frac{\bar{y}^2}{\bar{x}^2}\right] &= \frac{\mu_y^2}{\mu_x^2} + \frac{\text{Var}(\bar{y})}{\mu_y} + 3\text{Var}(\bar{x})\frac{\mu_y^2}{\mu_x^4} - 4\text{Cov}(\bar{x}, \bar{y})\frac{\mu_y}{\mu_x^3} \\ &\quad + \sum_{i,j: i+j \geq 3} \mathbb{E}\left[(\bar{x} - \mu_x)^i (\bar{y} - \mu_y)^j\right] \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y^2}{\mu_x^2}. \end{aligned} \quad (13)$$

Assuming $(x_1, y_1), \dots, (x_n, y_n) \sim \mathbb{P}_{xy}$ are i.i.d. further simplifies the terms, like in the following. Markov's inequality does not require this assumption, so a violation does not invalidate our approach. Then, it holds that

$$\text{Var}(\bar{x}) = \frac{1}{n} \text{Var}(x), \quad \text{Var}(\bar{y}) = \frac{1}{n} \text{Var}(y), \quad \text{Cov}(\bar{x}, \bar{y}) = \frac{1}{n} \text{Cov}(x, y). \quad (14)$$

Further, for all $a = 1, \dots, n$ let $z_{k,a} = x_a$ and $\mu_{z_k} = \mu_x$ if $1 \leq k \leq i$, and $z_{k,a} = y_a$ and $\mu_{z_k} = \mu_y$ if $i < k \leq m := i + j$. Then

$$\begin{aligned}
 & \mathbb{E} \left[(\bar{x} - \mu_x)^i (\bar{y} - \mu_y)^j \right] \\
 &= \frac{1}{n^{i+j}} \mathbb{E} \left[\left(\sum_{a=1}^n x_a - \mu_x \right)^i \left(\sum_{a=1}^n y_a - \mu_y \right)^j \right] \\
 &= \frac{1}{n^m} \mathbb{E} \left[\prod_{k=1}^m \left(\sum_{a=1}^n z_{k,a} - \mu_{z_k} \right) \right] \\
 &= \frac{1}{n^m} \sum_{l=1}^m \sum_{a_l=1}^n \mathbb{E} \left[\prod_{k=1}^m (z_{k,a_k} - \mu_{z_k}) \right]
 \end{aligned} \tag{15}$$

For all a_k holds that $\mathbb{E} \left[\prod_{k=1}^m (z_{k,a_k} - \mu_{z_k}) \right] = 0$ if there exists any non-duplicate index value, due to independence. It follows that we can reduce the number of indices by at least half, which reduces the number of addends by a polynomial:

$$\begin{aligned}
 & \frac{1}{n^m} \sum_{l=1}^m \sum_{a_l=1}^n \mathbb{E} \left[\underbrace{\prod_{k=1}^m (z_{k,a_k} - \mu_{z_k})}_{n^m \text{ addends}} \right] \\
 &= \frac{1}{n^m} \sum_{l=1}^{\lfloor m/2 \rfloor} \sum_{a_l=1}^n \mathbb{E} \left[\underbrace{\prod_{k=1}^m (z_{k,a_k} - \mu_{z_k})}_{n^{\lfloor m/2 \rfloor} \text{ addends}} \right] \\
 &= \frac{1}{n^{\lceil m/2 \rceil}} \underbrace{\frac{1}{n^{\lfloor m/2 \rfloor}} \sum_{l=1}^{\lfloor m/2 \rfloor} \sum_{a_l=1}^n \mathbb{E} \left[\prod_{k=1}^m (z_{k,a_k} - \mu_{z_k}) \right]}_{=: C_{ij}}.
 \end{aligned} \tag{16}$$

Note that $C_{ij} \in [-B_m, B_m]$ with $B_m := \max_{\{i,j=0,\dots,m \mid i+j \leq m\}} \left| \mathbb{E} \left[(x - \mu_x)^i (y - \mu_y)^j \right] \right|$, therefore, the convergence rate depends not only on the data size n but also on how the moments grow with m .

Using Eqn. 14 and Eqn. 16 gives

$$\begin{aligned}
 \mathbb{E} \left[\frac{\bar{y}^2}{\bar{x}^2} \right] &= \frac{\mu_y^2}{\mu_x^2} + \frac{\text{Var}(y)}{n\mu_y} + 3 \text{Var}(x) \frac{\mu_y^2}{n\mu_x^4} - 4 \text{Cov}(x, y) \frac{\mu_y}{n\mu_x^3} \\
 &+ \sum_{i,j: i+j \geq 3} \frac{1}{n^{\lceil (i+j)/2 \rceil}} C_{ij} \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y^2}{\mu_x^2}.
 \end{aligned} \tag{17}$$

Similarly, we use Taylor expansion for $\frac{\bar{y}}{\bar{x}}$ around $\frac{\mu_y}{\mu_x}$ to get

$$\begin{aligned}
 \frac{\bar{y}}{\bar{x}} &= \frac{\mu_y}{\mu_x} + (\bar{y} - \mu_y) \frac{1}{\mu_x} - (\bar{x} - \mu_x) \frac{\mu_y}{\mu_x^2} \\
 &+ 0 + (\bar{x} - \mu_x)^2 \frac{\mu_y}{\mu_x^3} - (\bar{y} - \mu_y) (\bar{x} - \mu_x) \frac{1}{\mu_x^2} \\
 &+ \sum_{i,j: i+j \geq 3} (\bar{y} - \mu_y)^i (\bar{x} - \mu_x)^j \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y}{\mu_x},
 \end{aligned} \tag{18}$$

which results in

$$\begin{aligned}
\frac{\mu_y}{\mu_x} \mathbb{E} \left[\frac{\bar{y}}{\bar{x}} \right] &= \frac{\mu_y^2}{\mu_x^2} + \text{Var}(\bar{x}) \frac{\mu_y^2}{\mu_x^4} - \text{Cov}(\bar{y}, \bar{x}) \frac{\mu_y}{\mu_x^3} \\
&\quad + \sum_{i,j: i+j \geq 3} \mathbb{E} \left[(\bar{y} - \mu_y)^i (\bar{x} - \mu_x)^j \right] \frac{\mu_y}{\mu_x} \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y}{\mu_x} \\
&= \frac{\mu_y^2}{\mu_x^2} + \text{Var}(x) \frac{\mu_y^2}{n \mu_x^4} - \text{Cov}(x, y) \frac{\mu_y}{n \mu_x^3} \\
&\quad + \sum_{i,j: i+j \geq 3} \frac{1}{n^{\lceil (i+j)/2 \rceil}} C_{ij} \frac{\mu_y}{\mu_x} \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y}{\mu_x}.
\end{aligned} \tag{19}$$

Inserting Eqn. 17 and Eqn. 19 into Eqn. 11 results in

$$\begin{aligned}
\text{SE}_{\hat{r}} &= 2 \frac{\mu_y^2}{\mu_x^2} + \frac{\text{Var}(y)}{n \mu_x} + 3 \text{Var}(x) \frac{\mu_y^2}{n \mu_x^4} - 4 \text{Cov}(x, y) \frac{\mu_y}{n \mu_x^3} \\
&\quad + \sum_{i,j: i+j \geq 3} \frac{1}{n^{\lceil (i+j)/2 \rceil}} C_{ij} \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y^2}{\mu_x^2} \\
&\quad - 2 \left(\frac{\mu_y^2}{\mu_x^2} + \text{Var}(x) \frac{\mu_y^2}{n \mu_x^4} - \text{Cov}(x, y) \frac{\mu_y}{n \mu_x^3} \right. \\
&\quad \left. + \sum_{i,j: i+j \geq 3} \frac{1}{n^{\lceil (i+j)/2 \rceil}} C_{ij} \frac{\mu_y}{\mu_x} \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y}{\mu_x} \right) \\
&= \frac{1}{n} \left(\frac{\text{Var}(y)}{\mu_x} + \text{Var}(x) \frac{\mu_y^2}{\mu_x^4} - 2 \text{Cov}(x, y) \frac{\mu_y}{\mu_x^3} \right) \\
&\quad + \underbrace{\sum_{i,j: i+j \geq 3} \frac{1}{n^{\lceil (i+j)/2 \rceil}} C_{ij} \left(\frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y^2}{\mu_x^2} - \frac{2 \mu_y}{\mu_x} \frac{\partial^{i+j}}{(\partial \mu_x)^i (\partial \mu_y)^j} \frac{\mu_y}{\mu_x} \right)}_{\in O\left(\frac{1}{n^2}\right)}.
\end{aligned} \tag{20}$$

Consequently, we may estimate $\text{SE}_{\hat{r}}$ via

$$\widehat{\text{SE}}_{\hat{r}} := \frac{1}{n} \left(\frac{\hat{\mu}_y \hat{\sigma}_x^2}{\hat{\mu}_x^4} + \frac{\hat{\sigma}_y^2}{\hat{\mu}_x} - 2 \frac{\hat{\mu}_y \hat{\sigma}_{xy}}{\hat{\mu}_x^3} \right), \tag{21}$$

which is consistent since the estimators $\hat{\mu}_y = \frac{1}{n} \sum_i y_i$, $\hat{\mu}_x = \frac{1}{n} \sum_i x_i$, $\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_i (y_i - \hat{\mu}_y)^2$, $\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_i (x_i - \hat{\mu}_x)^2$, and $\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$ are consistent as well.

B.2 VOLUME TO RATIO CONFIDENCE INTERVALS

Note that if $a \notin [b, c] \subseteq \mathbb{R}_{>0}$ then $\frac{1}{a} \notin [\frac{1}{c}, \frac{1}{b}]$ since $x \mapsto \frac{1}{x}$ is strictly negative monotone. We also make use of the subadditivity of probability measures (Resnick, 2003) given by

$$\mathbb{P} \left(\bigcup_i A_i \right) \leq \sum_i \mathbb{P}(A_i). \tag{22}$$

This is also known as Boole's inequality. Denote the random variable z representing pixel inputs of image instance I . Let $q_{T,\alpha}$ and $q_{N,\alpha}$ be empirically determined on a validation set as the $1 - \alpha$

quantile of the image-wise calibration errors for g_T and g_N . Then, for $\alpha \in [0, 1]$ it holds

$$\begin{aligned}
\alpha &= \frac{\alpha}{2} + \frac{\alpha}{2} \\
&\geq \mathbb{P}(\text{CE}_{N,I} \geq q_{N,\alpha/2}) + \mathbb{P}(\text{CE}_{T,I} \geq q_{T,\alpha/2}) \\
&\geq \mathbb{P}(|\mathbb{E}[Y_N | I] - \mathbb{E}[g_N(z) | I]| \geq q_{N,\alpha/2}) + \mathbb{P}(|\mathbb{E}[Y_T | I] - \mathbb{E}[g_T(z) | I]| \geq q_{T,\alpha/2}) \\
&\geq \mathbb{P}(|\mathbb{E}[Y_N | I] - \mathbb{E}[g_N(z) | I]| \geq q_{N,\alpha/2} \vee |\mathbb{E}[Y_T | I] - \mathbb{E}[g_T(z) | I]| \geq q_{T,\alpha/2}) \\
&= \mathbb{P}\left(\mathbb{E}[Y_N | I] \notin [\mathbb{E}[g_N(z) | I] - q_{N,\alpha}, \mathbb{E}[g_N(z) | I] + q_{N,\alpha}] \right. \\
&\quad \left. \vee \mathbb{E}[Y_T | I] \notin [\mathbb{E}[g_T(z) | I] - q_{T,\alpha}, \mathbb{E}[g_T(z) | I] + q_{T,\alpha}]\right) \\
&= \mathbb{P}\left(\mathbb{E}[Y_N | I] \notin [\mathbb{E}[g_N(z) | I] - q_{N,\alpha}, \mathbb{E}[g_N(z) | I] + q_{N,\alpha}] \right. \\
&\quad \left. \vee \frac{1}{\mathbb{E}[Y_T | I]} \notin \left[\frac{1}{\mathbb{E}[g_T(z) | I] + q_{T,\alpha}}, \frac{1}{\mathbb{E}[g_T(z) | I] - q_{T,\alpha}}\right]\right) \\
&\geq \mathbb{P}\left(\frac{\mathbb{E}[Y_N | I]}{\mathbb{E}[Y_T | I]} \notin \left[\frac{\mathbb{E}[g_N(z) | I] - q_{N,\alpha}}{\mathbb{E}[g_T(z) | I] + q_{T,\alpha}}, \frac{\mathbb{E}[g_N(z) | I] + q_{N,\alpha}}{\mathbb{E}[g_T(z) | I] - q_{T,\alpha}}\right]\right).
\end{aligned} \tag{23}$$

It follows that for confidence level $1 - \alpha$ that

$$\frac{\mathbb{E}[Y_N | I]}{\mathbb{E}[Y_T | I]} \in \left[\frac{\mathbb{E}[g_N(z) | I] - q_{N,\alpha}}{\mathbb{E}[g_T(z) | I] + q_{T,\alpha}}, \frac{\mathbb{E}[g_N(z) | I] + q_{N,\alpha}}{\mathbb{E}[g_T(z) | I] - q_{T,\alpha}}\right] \tag{24}$$

Given the previous equation, it further holds that

$$\begin{aligned}
\delta + \alpha &\geq \\
&\geq \mathbb{P}\left(\frac{\mathbb{E}[Y_N | I]}{\mathbb{E}[Y_T | I]} \notin \left[\frac{\mathbb{E}[g_N(z) | I]}{\mathbb{E}[g_T(z) | I]} - \epsilon_{l,\delta}, \frac{\mathbb{E}[g_N(z) | I]}{\mathbb{E}[g_T(z) | I]} + \epsilon_{u,\delta}\right]\right) \\
&\quad + \mathbb{P}\left(\frac{\mathbb{E}[g_N(z) | I]}{\mathbb{E}[g_T(z) | I]} \notin \left[\frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} - \beta_{r,\alpha}, \frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} + \beta_{r,\alpha}\right]\right) \\
&\geq \mathbb{P}\left(\frac{\mathbb{E}[Y_N | I]}{\mathbb{E}[Y_T | I]} \notin \left[\frac{\mathbb{E}[g_N(z) | I]}{\mathbb{E}[g_T(z) | I]} - \epsilon_{l,\delta}, \frac{\mathbb{E}[g_N(z) | I]}{\mathbb{E}[g_T(z) | I]} + \epsilon_{u,\delta}\right]\right) \\
&\quad \vee \frac{\mathbb{E}[g_N(z) | I]}{\mathbb{E}[g_T(z) | I]} \notin \left[\frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} - \beta_{r,\alpha}, \frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} + \beta_{r,\alpha}\right] \\
&\geq \mathbb{P}\left(\frac{\mathbb{E}[Y_N | I]}{\mathbb{E}[Y_T | I]} \notin \left[\frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} - \epsilon_{l,\delta} - \beta_{r,\alpha}, \frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} + \epsilon_{u,\delta} + \beta_{r,\alpha}\right]\right).
\end{aligned} \tag{25}$$

From this follows that with at least probability $1 - \alpha - \delta$ that

$$\frac{\mathbb{E}[Y_N | I]}{\mathbb{E}[Y_T | I]} \in \left[\frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} - \epsilon_{l,\delta} - \beta_{r,\alpha}, \frac{\sum_i g_N(z_{i,I})}{\sum_i g_T(z_{i,I})} + \epsilon_{u,\delta} + \beta_{r,\alpha}\right]. \tag{26}$$

B.3 DEBIASED RATIO ESTIMATION

The naive ratio estimator is biased due to the limited number of samples. Here we extend Popordanoska et al. (2022) to derive a debiased ratio estimator to $\mathcal{O}(n^{-2})$. Firstly, the naive estimator is:

$$\hat{r} = \frac{\bar{y}}{\bar{x}} = \frac{\mu_y}{\mu_x} \left(\frac{\bar{y}}{\mu_y}\right) \left(\frac{\bar{x}}{\mu_x}\right)^{-1} = \frac{\mu_y}{\mu_x} \left(1 + \frac{\bar{y} - \mu_y}{\mu_y}\right) \left(1 + \frac{\bar{x} - \mu_x}{\mu_x}\right)^{-1}. \tag{27}$$

Then we expand $\left(1 + \frac{\bar{x} - \mu_x}{\mu_x}\right)^{-1}$ in Taylor series:

$$\begin{aligned} \hat{r} = \frac{\mu_y}{\mu_x} & \left(1 + \frac{(\bar{y} - \mu_y)}{\mu_y} - \frac{(\bar{x} - \mu_x)}{\mu_x} - \frac{(\bar{x} - \mu_x)(\bar{y} - \mu_y)}{\mu_y \mu_x} + \frac{(\bar{x} - \mu_x)^2}{\mu_x^2} \right. \\ & \left. + \frac{(\bar{x} - \mu_x)^2(\bar{y} - \mu_y)}{\mu_x^2 \mu_y} - \frac{(\bar{x} - \mu_x)^3}{\mu_x^3} - \frac{(\bar{x} - \mu_x)^3(\bar{y} - \mu_y)}{\mu_x^3 \mu_y} + \frac{(\bar{x} - \mu_x)^4}{\mu_x^4} \right) + \mathcal{O}(n^{-2.5}) \end{aligned} \quad (28)$$

The bias of $\hat{r}$ defined by $\mathbb{E}[\hat{r}] - r$ is written as:

$$\text{Bias}_r = \frac{\mu_y}{\mu_x} \left(\frac{1}{n} \left(\frac{\text{Var}(x)}{\mu_x^2} - \frac{\text{Cov}(x, y)}{\mu_x \mu_y} \right) + \frac{1}{n^2} \left(\frac{(\text{Cov}(x^2, y) - 2\mu_x \text{Cov}(x, y))}{\mu_x^2 \mu_y} \right. \right. \quad (29)$$

$$\left. - \frac{(\text{Cov}(x^2, x) - 2\mu_x \text{Var}(x))}{\mu_x^3} - \frac{3 \text{Var}(x) \text{Cov}(x, y)}{\mu_x^3 \mu_y} + \frac{3 \text{Var}(x)^2}{\mu_x^4} \right) \quad (30)$$

And a second-order debiased estimator is defined by $r_{corr,2} := \hat{r} - \text{Bias}_r$:

$$r_{corr,2} = \hat{r} - \frac{\mu_y}{\mu_x} \left(\frac{1}{n} \left(\frac{\text{Var}(x)}{\mu_x^2} - \frac{\text{Cov}(x, y)}{\mu_x \mu_y} \right) + \frac{1}{n^2} \left(\frac{(\text{Cov}(x^2, y) - 2\mu_x \text{Cov}(x, y))}{\mu_x^2 \mu_y} \right. \right. \quad (31)$$

$$\left. - \frac{(\text{Cov}(x^2, x) - 2\mu_x \text{Var}(x))}{\mu_x^3} - \frac{3 \text{Var}(x) \text{Cov}(x, y)}{\mu_x^3 \mu_y} + \frac{3 \text{Var}(x)^2}{\mu_x^4} \right) \quad (32)$$

Finally, we use plug-in estimators for empirical estimation:

$$\begin{aligned} \hat{r}_{corr,2} := \frac{\hat{\mu}_y}{\hat{\mu}_x} & \left(1 - \frac{1}{n} \left(r_b^* - r_a^* \right) - \frac{1}{n^2} \left(\frac{(\widehat{\text{Cov}}(x^2, y) - 2\hat{\mu}_x \widehat{\text{Cov}}(x, y))}{\hat{\mu}_x^2 \hat{\mu}_y} \right. \right. \\ & \left. - \frac{(\widehat{\text{Cov}}(x^2, x) - 2\hat{\mu}_x \widehat{\text{Var}}(x))}{\hat{\mu}_x^3} - \frac{3\widehat{\text{Var}}(x) \widehat{\text{Cov}}(x, y)}{\hat{\mu}_x^3 \hat{\mu}_y} + \frac{3\widehat{\text{Var}}(x)^2}{\hat{\mu}_x^4} \right) \end{aligned} \quad (33)$$

$$\begin{aligned} r_a^* = \underbrace{\frac{\widehat{\text{Cov}}(x, y)}{\widehat{\mu}_x \widehat{\mu}_y}}_{=r_a} & \left(1 + \frac{1}{(n-1)} \left(\frac{\widehat{\mu}_y \widehat{\text{Cov}}(x^2, y) + \widehat{\mu}_x \widehat{\text{Cov}}(y^2, x)}{\widehat{\text{Cov}}(x, y) \widehat{\mu}_x \widehat{\mu}_y} - 4 \right) \right. \\ & \left. - \frac{1}{(n-1)} \left(\frac{\widehat{\text{Var}}(x)}{\widehat{\mu}_x^2} + \frac{\widehat{\text{Var}}(y)}{\widehat{\mu}_y^2} + 2 \frac{\widehat{\text{Cov}}(x, y)}{\widehat{\mu}_x \widehat{\mu}_y} \right) \right) \end{aligned} \quad (34)$$

$$r_b^* = \underbrace{\frac{\widehat{\text{Var}}(x)}{\widehat{\mu}_x^2}}_{=r_b} \left(1 + \frac{4}{(n-1)} \left(\frac{\frac{1}{2} \widehat{\text{Cov}}(x^2, x)}{\widehat{\mu}_x \widehat{\text{Var}}(x)} - 1 \right) - \frac{4}{(n-1)} \frac{\widehat{\text{Var}}(x)}{\widehat{\mu}_x^2} \right). \quad (35)$$

C LLM USAGE

We use ChatGPT (OpenAI, 2025) for polishing the writing of this paper, including improving grammar and clarity. No part of the paper was generated solely by LLM. All technical content, ideas, and experimental results were created and validated by the authors.