Minimax Multi-Target Conformal Prediction with Applications to Imaging Inverse Problems

Anonymous authors
Paper under double-blind review

Abstract

In ill-posed imaging inverse problems, uncertainty quantification remains a fundamental challenge, especially in safety-critical applications. Recently, conformal prediction has been used to quantify the uncertainty that the inverse problem contributes to downstream tasks like image classification, image quality assessment, fat mass quantification, etc. While existing works handle only a scalar estimation target, practical applications often involve multiple targets. In response, we propose a minimax approach to multi-target conformal prediction that provides tight prediction intervals while ensuring joint marginal coverage. We then outline how our minimax approach can be applied to multi-metric blind image quality assessment, multi-task uncertainty quantification, and multi-round measurement acquisition. Finally, we numerically demonstrate the benefits of our minimax method, relative to existing multi-target conformal prediction methods, using magnetic resonance imaging (MRI) data.

1 Introduction

Imaging inverse problems (Bertero et al., 2021) span a wide array of tasks, such as denoising, inpainting, accelerated magnetic resonance imaging (MRI), limited-angle computed tomography, phase retrieval, and image-to-image translation. In such problems, the objective is to recover a true image x_0 from noisy, incomplete, or distorted measurements $y_0 = \mathcal{A}(x_0)$. These problems tend to be ill-posed, in that many distinct hypotheses of x_0 can explain the collected measurements y_0. When perfect recovery of x_0 is difficult or impossible, uncertainty quantification (UQ) is critical to safely using/interpreting a given reconstruction $\widehat{x}_0$, especially in high-stakes fields like science or medicine (Chu et al., 2020; Banerji et al., 2023).

The field of image recovery has evolved significantly over the decades, and most contemporary approaches are based on deep learning (DL) (Arridge et al., 2019). Quantitatively, recent DL-based methods outperform classical methods on average and, qualitatively, they produce reconstructions that are sharp and detailed (Ongie et al., 2020). When the inverse problem is highly ill-posed, classical methods tend to produce recoveries with recognizable visual artifacts, from which it is relatively easy to gauge uncertainty, and radiologists receive explicit training in this regard (Virmani et al., 2015). In contrast, DL-based methods can hallucinate, i.e., generate recoveries that are visually plausible but differ from the truth in clinically or scientifically important ways (Cohen et al., 2018; Belthangady & Royer, 2019; Hoffman et al., 2021; Muckley et al., 2021; Bhadra et al., 2021; Gottschling et al., 2023; Tivnan et al., 2024). This underscores the need for rigorous UQ, e.g., methods that provide statistical guarantees on estimates of x_0 or of some function $\mu(x_0)$.

For example, a recent line of work (Wen et al., 2024; Cheung et al., 2024) quantifies the imaging-induced uncertainty on downstream tasks such as pathology classification or fat-mass quantification. Defining the target z_0 as the output of the task applied to the (unknown) true image, they use conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) to construct prediction intervals $\mathcal{C}$ that are statistically guaranteed to contain the target. In a related line of work, Wen et al. (2025) provides statistical guarantees on the quality of the reconstructed image $\widehat{x}_0$ relative to the true image, where "quality" is defined according to an arbitrary full-reference-image-quality (FRIQ) metric like peak signal-to-noise ratio (PSNR) or structural similarity index measure (SSIM) (Wang et al., 2004). Defining the target as the FRIQ of $\widehat{x}_0$ relative to the (unknown) true x_0, they use conformal prediction to construct a bound on FRIQ that is statistically guaranteed.

While the above methods rigorously quantify the downstream impact of reconstruction uncertainty, they handle only a single target. In practice, one may want to consider multiple targets. For example, one may seek to identify multiple pathologies from a single recovery or to judge the quality of that recovery according to multiple metrics. Although multi-target conformal prediction methods have been proposed, they suffer from either limited interpretability (Messoudi et al., 2022; Feldman et al., 2023), a lack of guaranteed joint coverage (Messoudi et al., 2021; Teneggi et al., 2023; Park et al., 2024), or conservative prediction intervals (Messoudi et al., 2020; Diquigiovanni et al., 2022; Sampson & Chan, 2024), as we explain in the sequel.

We thus propose a new approach to multi-target conformal prediction. For problems with $K \geq 1$ targets, our goal is to ensure that the prediction interval $\mathcal{C}_k$ is not overly large for any target $k \in \{1, \ldots, K\}$ while guaranteeing that all prediction intervals simultaneously contain their corresponding targets $Z_{0,k}$ with a user-specified probability of $1 - \alpha$. Leveraging the fact that the "single-target coverage" $\Pr\{Z_{0,k} \in \mathcal{C}_k\}$ increases with the interval size $|\mathcal{C}_k|$, our method aims to minimize the maximum single-target coverage while obeying a joint coverage guarantee of the form $\Pr\{\cap_k Z_{0,k} \in \mathcal{C}_k\} \leq 1 - \alpha$. Our contributions are as follows:

1. We propose a novel multi-target conformal prediction approach based on minimax optimization.
2. We prove the statistical convergence of our method in the large-sample regime.
3. For inverse problems, we propose a multi-round measurement acquisition scheme with performance guarantees on the final round.
4. We numerically compare our proposed method to several existing multi-target conformal prediction methods on four accelerated-MRI problems.

2 Background

2.1 Single-target conformal prediction

Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) is a general framework that enables one to construct uncertainty intervals for any black-box predictor with certain statistical guarantees. Importantly, it does not require any distributional assumptions on the data other than exchangeability, which allows for adoption in a broad range of applications. In this section, we briefly review the basics of conformal prediction, and in particular the computationally-efficient version known as split conformal prediction (Papadopoulos et al., 2002; Lei et al., 2018).

Suppose that we have a black-box model $h : \mathcal{U} \to \mathbb{R}$ that predicts a target $z_0 \in \mathbb{R}$ from features $u_0 \in \mathcal{U}$. The prediction $\widehat{z}_0 = h(u_0)$ may or may not be close to the true target z_0, but one can use conformal prediction to compute a prediction interval $\mathcal{C}_\lambda(\widehat{z}_0) \subset \mathbb{R}$ that contains z_0 with high probability. To compute this interval, conformal prediction uses a dataset $\{(u_i, z_i)\}_{i=1}^n$ of feature–target pairs distinct from those used to train $h(\cdot)$. This dataset is converted to a calibration set $d_{\mathsf{cal}} \triangleq \{(\widehat{z}_i, z_i)\}_{i=1}^n$ using $\widehat{z}_i = h(u_i)$, and d_{cal} is used to find a $\widehat{\lambda}(d_{\mathsf{cal}})$ satisfying the marginal coverage guarantee (Lei & Wasserman, 2014)

$$\Pr\left\{Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{Z}_0)\right\} \geq 1 - \alpha, \tag{1}$$

where α is a user-chosen error rate. Here and in the sequel, we use capital letters to denote random variables and lower-case letters to denote their realizations. In words, (1) guarantees that the unknown target Z_0 falls within the interval $\mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{Z}_0)$ with probability at least $1 - \alpha$ when averaged over the randomness in the test data $(Z_0, \widehat{Z}_0)$ and calibration data D_{cal}.

The process of computing $\widehat{\lambda}(d_{\mathsf{cal}})$ is known as calibration. To calibrate, one first defines a nonconformity score $s(\widehat{z}_i, z_i)$. The choice of the nonconformity score function is quite flexible; it requires only that the score is higher when there is a worse agreement between z_i and $\widehat{z}_i$. Common choices include the absolute residual, locally-weighted residual (Lei et al., 2018), and conformalized quantile regression methods (Romano et al., 2019). The nonconformity score $s_i = s(\widehat{z}_i, z_i)$ is computed for each sample pair $(\widehat{z}_i, z_i)$ in the calibration set

d_{cal}, and $\widehat{\lambda}(d_{\mathsf{cal}})$ is chosen as

$$\widehat{\lambda}(d_{\mathsf{cal}}) \triangleq \operatorname{EmpQuant}\Big(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}; s_1, \ldots, s_n\Big), \tag{2}$$

which is a slightly more conservative quantile than the $1-\alpha$ quantile. With $\widehat{\lambda}(d_{\mathsf{cal}})$ computed, the prediction interval for the i-th sample is simply defined as

$$\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\widehat{z}_i) = \{z : s(\widehat{z}_i, z) \leq \widehat{\lambda}(d_{\mathsf{cal}})\}. \tag{3}$$

Following this design, the marginal coverage guarantee (1) holds when $(\widehat{Z}_0, Z_0), (\widehat{Z}_1, Z_1), \ldots, (\widehat{Z}_n, Z_n)$ are statistically exchangeable (Vovk et al., 2005), a weaker condition than i.i.d. Under the additional assumption that the nonconformity scores $S_1, \ldots, S_n$ are almost surely distinct, the coverage can also be upper bounded (Romano et al., 2019) by

$$\Pr\big\{Z_0 \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}})}(\widehat{Z}_0)\big\} \leq 1 - \alpha + \frac{1}{n+1}. \tag{4}$$

2.2 Application to imaging inverse problems

In imaging inverse problems, conformal prediction has emerged as a tool to quantify the uncertainty in image recovery. Several approaches (Angelopoulos et al., 2022; Horwitz & Hoshen, 2022; Teneggi et al., 2023; Kutiel et al., 2023; Narnhofer et al., 2024) use conformal prediction to construct, for each individual pixel, an interval that is guaranteed to contain the true pixel value with high probability. For quantifying multi-pixel uncertainty, Belhasin et al. (2023) propose to compute conformal intervals on the principal components of the posterior covariance matrix, and Sankaranarayanan et al. (2022) construct conformal intervals for semantic attributes in the latent space of a disentangled generative adversarial network.

Although these notions of uncertainty are interesting to consider, they don't directly quantify the impact of recovery errors on downstream imaging tasks such as image classification, image quality assessment, and quantitative imaging. Consequently, task-based image uncertainty methods like (Wen et al., 2024; Cheung et al., 2024; Wen et al., 2025) have been proposed. We now briefly review these methods using a unified notational framework.

Both Wen et al. (2024) and Cheung et al. (2024) quantify the uncertainty in estimating $\mu(x_0) \in \mathbb{R}$ given the measurements y_0. In Wen et al. (2024) $\mu(\cdot)$ is a soft-output classifier, and in Cheung et al. (2024) it is a fat-mass quantifier, but in either case the target is set at $z_0 = \mu(x_0)$. Assuming access to an approximate posterior sampler $g(\cdot,\cdot)$ that generates c samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^c$ per measurement vector y_i via $\widetilde{x}_i^{(j)} = g(y_i, \widetilde{v}_i^{(j)})$ using i.i.d code vectors $\widetilde{v}_i^{(j)} \sim \mathcal{N}(0, I)$, the prediction is computed as

$$\widehat{z}_i = [\widehat{z}_i^{(1)}, \ldots, \widehat{z}_i^{(c)}]^\top = [\mu(\widetilde{x}_i^{(1)}), \ldots, \mu(\widetilde{x}_i^{(c)})]^\top \triangleq h(u_i) \in \mathbb{R}^c, \tag{5}$$

where $h(u_i)$ and $u_i \triangleq [\widetilde{x}_i^{(1)}, \ldots, \widetilde{x}_i^{(c)}]^\top$ follow the notation introduced just before (1).

Several nonconformity scores can be used, but here we describe only the conformalized quantile regression (CQR) method of Romano et al. (2019). In CQR, the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ empirical quantiles are computed as

$$\widehat{q}_{\frac{\alpha}{2}}(\widehat{z}_i) = \operatorname{EmpQuant}\Big(\frac{\alpha}{2}; \widehat{z}_i^{(1)}, \ldots, \widehat{z}_i^{(c)}\Big) \quad \text{and} \quad \widehat{q}_{1-\frac{\alpha}{2}}(\widehat{z}_i) = \operatorname{EmpQuant}\Big(1-\frac{\alpha}{2}; \widehat{z}_i^{(1)}, \ldots, \widehat{z}_i^{(c)}\Big), \tag{6}$$

respectively, and the nonconformity score is defined as

$$s(\widehat{z}_i, z_i) = \max\{\widehat{q}_{\frac{\alpha}{2}}(\widehat{z}_i) - z_i, z_i - \widehat{q}_{1-\frac{\alpha}{2}}(\widehat{z}_i)\}, \tag{7}$$

after which the prediction interval $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}})}(\widehat{z}_i)$ is constructed as in (3). Because this prediction interval changes with y_i, it is said to be "adaptive" (Lei et al., 2018). In any case, it satisfies the marginal coverage guarantee in (1) when $(\widehat{Z}_0, Z_0), (\widehat{Z}_1, Z_1), \ldots, (\widehat{Z}_n, Z_n)$ are statistically exchangeable.

In related work, Wen et al. (2025) seek to estimate the FRIQ (e.g., PSNR, SSIM, etc.) $m(\widehat{x}_0, x_0)$ of an image recovery $\widehat{x}_0 = f(y_0)$ relative to the true image x_0 when given access to measurements y_0 but not x_0 itself. To do so, they set the target as $z_0 = m(\widehat{x}_0, x_0)$ and use an approximate posterior sampler that generates c samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^c$ per measurement vector y_i to compute the prediction

$$\widehat{z}_i = [m(\widehat{x}_i, \widetilde{x}_i^{(1)}), \dots, m(\widehat{x}_i, \widetilde{x}_i^{(c)})]^\top \triangleq h(u_i) \in \mathbb{R}^c, \tag{8}$$

where $h(u_i)$ and $u_i \triangleq [\widetilde{x}_i^{(1)}, \dots, \widetilde{x}_i^{(c)}, \widehat{x}_i]^\top$ follow the notation introduced just before (1). Wen et al. (2025) then used empirical quantiles to construct a one-sided prediction interval $\mathcal{C}_\lambda(\cdot)$ to either lower- or upper-bound the FRIQ, as appropriate. Note that $m(\widehat{x}_i, \cdot)$ can be viewed as a recovery-conditioned task. To maintain consistency with other task-based approaches, when discussing FRIQ estimation in the sequel, we construct two-sided intervals using (3) with the nonconformity score from (7).

2.3 Multi-target conformal prediction

As discussed in Section 1, one may be interested in conformal prediction of several targets, which we combine into a multi-dimensional target vector $[z_{i,1}, \dots, z_{i,K}] = z_i \in \mathbb{R}^K$. We focus on the case where one is given a prediction $\widehat{z}_0 \in \mathbb{R}^K$ of unknown test $z_0 \in \mathbb{R}^K$, along with a calibration set $d_{\mathsf{cal}} = \{(\widehat{z}_i, z_i)\}_{i=1}^n$, and the goal is to compute K prediction intervals $\{\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}),k}(\widehat{z}_0)\}_{k=1}^K$ that satisfy the joint marginal coverage guarantee

$$\Pr\big\{ \cap_{k=1}^K Z_{0,k} \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}}),k}(\widehat{Z}_0)\big\} \geq 1-\alpha. \tag{9}$$

This guarantee ensures that all target components simultaneously lie within their respective prediction intervals with a probability at least $1-\alpha$ over the randomness in the calibration and test data.

Variations on (9) are possible, such as minimizing a risk that allows a fraction of the target components to lie outside their prediction intervals (Teneggi et al., 2023), but we focus on the stricter requirement (9). Likewise, while (9) can be interpreted as constructing a hyper-rectangle in $\mathbb{R}^K$ that contains Z_0 with high probability, it is possible to construct non-rectangular regions, such as ellipsoidal regions (Messoudi et al., 2022) or more complicated regions defined by the outputs of a a conditional variational auto-encoder (Feldman et al., 2023). Although these non-rectangular regions can give smaller uncertainty volumes, they are less interpretable, since the uncertainty interval on one target component will depend on the values of the other target components.

Inspired by (9), several approaches have been proposed to construct prediction intervals. Messoudi et al. (2020) assume that the nonconformity score components are statistically independent, so that when the components are individually calibrated to yield an error-rate of α_1, the joint error-rate α will equal $1-(1-\alpha_1)^K$. This allows setting $\alpha_1 = 1-(1-\alpha)^{1/K}$ to meet a desired joint error-rate of α. However, the independence assumption may not hold in practice, where one could encounter K dependent score components that yield a joint error-rate $> 1-(1-\alpha_1)^K$, in which case the joint coverage guarantee (9) would be violated. But even when the joint coverage holds, we show in Section 5 that the prediction intervals from Messoudi et al. (2020) are overly conservative.

Another line of work (Messoudi et al., 2021; Park et al., 2024) uses copulas (Nelsen, 2006) to model the statistical dependency between the score components $\{s_{i,k}\}_{k=1}^K$. There, the calibration data is used to compute empirical estimates of the copula and the marginal distributions, from which the joint cumulative distribution function (CDF) of the nonconformity scores, $F_S(\cdot)$, is approximated as $\widehat{F}_S(\cdot)$. A $\widehat{\lambda}(d_{\mathsf{cal}}) \in \mathbb{R}^K$ is then computed that satisfies $\widehat{F}_S(\widehat{\lambda}(d_{\mathsf{cal}})) \geq 1-\alpha$. However, the quality of the CDF approximation depends on the choice of the copula, and there is no guarantee of satisfying a coverage guarantee like (9) with a finite calibration set.

As an alternative, Diquigiovanni et al. (2022) and Sampson & Chan (2024) propose to individually compute the nonconformity score components as $s_{i,k} = s_k(\widehat{z}_{i,k}, z_{i,k})$ for each k and combine them into a single score

$$s_i = \max\{s_{i,1}, \dots, s_{i,K}\}. \tag{10}$$

Using calibration $\{s_i\}_{i=1}^n$ with (2) and extending (3) to component-wise intervals

$$\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}),k}(\widehat{z}_i) = \{z : s_k(\widehat{z}_{i,k}, z) \leq \widehat{\lambda}(d_{\mathsf{cal}})\}, \quad k = 1, \dots, K, \tag{11}$$

the arguments from Vovk et al. (2005) imply that the joint coverage guarantee (9) and upper bound (4) both hold under the usual exchangeability assumption. However, Sampson & Chan (2024) note that this approach can disproportionally favor the target components with larger nonconformity scores, causing the prediction intervals of the other components to be overly conservative. To mitigate this issue, they propose to scale the nonconformity scores to a common range. To do so, they first train a pair of quantile regressors $\widehat{q}_{\frac{\alpha}{2},k}(u_i)$ and $\widehat{q}_{1-\frac{\alpha}{2},k}(u_i)$, which estimate the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantile of $Z_{i,k}$, respectively, for each $k \in \{1,\dots,K\}$, and then form the scaled nonconformity scores

$$s_{i,k} = \max\{\widehat{q}_{\frac{\alpha}{2},k}(\widehat{z}_i) - z_{i,k}, z_{i,k} - \widehat{q}_{1-\frac{\alpha}{2},k}(\widehat{z}_i)\} \underbrace{\frac{\widehat{q}_{1-\frac{\alpha}{2},1}(\widehat{z}_i) - \widehat{q}_{\frac{\alpha}{2},1}(\widehat{z}_i)}{\widehat{q}_{1-\frac{\alpha}{2},k}(\widehat{z}_i) - \widehat{q}_{\frac{\alpha}{2},k}(\widehat{z}_i)}}_{\text{Scale relative to 1st target}}, \tag{12}$$

where $\widehat{z}_i = u_i$. This helps to balance the marginal coverages $\Pr\{Z_{0,k} \in \mathcal{C}_{\widehat{\lambda}(D_{\mathsf{cal}}),k}(\widehat{Z}_0)\}$ across $k \in \{1,\dots,K\}$ while ensuring the joint coverage guarantee (9) and upper bound (4). In Section 3, we show how to obtain a better balance and thus tighter prediction intervals.

3 Minimax multi-target conformal prediction

In this section, we formulate a new approach to multi-target conformal prediction that is based on minimax optimization. First we describe the optimization problem in the context of random variables, then we describe our conformal method in the finite sample case, and finally we prove that the finite-sample method converges to a solution of the original minimax optimization as the number of samples grows to infinity.

3.1 Random variable perspective

To build intuition, we first consider the design of prediction sets when the targets and predictions are modeled as random variables $Z = [Z_1,\dots,Z_K] \in \mathbb{R}^K$ and $\widehat{Z} = [\widehat{Z}_1,\dots,\widehat{Z}_K]$, respectively. For the kth component, suppose that the nonconformity score function is $s_k(\cdot,\cdot)$ and the prediction set is constructed as $\mathcal{C}_{\widehat{\zeta}_k}(\widehat{Z}_k) \triangleq \{z : s_k(\widehat{Z}_k, z) \le \widehat{\zeta}_k\}$, where $\widehat{\zeta}_k$ is a design variable. Then the "single-target" coverage of the kth component will be

$$\Pr\big\{Z_k \in \mathcal{C}_{\widehat{\zeta}_k}(\widehat{Z}_k)\big\} = \Pr\big\{Z_k \in \{z : s_k(\widehat{Z}_k, z) \le \widehat{\zeta}_k\}\big\} = \Pr\big\{s_k(\widehat{Z}_k, Z_k) \le \widehat{\zeta}_k\big\}. \tag{13}$$

Using $S_k \triangleq s_k(\widehat{Z}_k, Z_k)$, we can write the single-target coverage more succinctly as

$$\Pr\big\{S_k \le \widehat{\zeta}_k\big\} = F_{S_k}(\widehat{\zeta}_k), \tag{14}$$

where $F_{S_k}(\widehat{\zeta}_k)$ is the CDF of S_k evaluated at $\widehat{\zeta}_k$. Similarly, the joint coverage of all K components will be

$$\Pr\big\{\cap_{k=1}^K Z_k \in \mathcal{C}_{\widehat{\zeta}_k}(\widehat{Z}_k)\big\} = \Pr\big\{\cap_{k=1}^K S_k \le \widehat{\zeta}_k\big\}. \tag{15}$$

For a given joint miscoverage rate of α, we'd like to find a tuple $(\widehat{\zeta}_1,\dots,\widehat{\zeta}_K)$ that ensures

$$\Pr\big\{\cap_{k=1}^K S_k \le \widehat{\zeta}_k\big\} \ge 1-\alpha. \tag{16}$$

But, in general, many $(\widehat{\zeta}_1,\dots,\widehat{\zeta}_K)$ will yield the same value of $\Pr\{\cap_{k=1}^K S_k \le \widehat{\zeta}_k\}$, and for some choices of $(\widehat{\zeta}_1,\dots,\widehat{\zeta}_K)$, a portion of the prediction intervals $\mathcal{C}_{\widehat{\zeta}_k}(\widehat{Z}_k)$ may be overly conservative. Since a larger $\Pr\{S_k \le \widehat{\zeta}_k\}$ generally corresponds to a larger prediction interval $\mathcal{C}_{\widehat{\zeta}_k}(\widehat{Z}_k)$ due to their monotonically non-decreasing relationship, we propose to design prediction intervals that minimize the maximum single-target coverage while ensuring joint coverage, i.e.,

$$(\widehat{\zeta}_1,\dots,\widehat{\zeta}_K) = \arg\min_{\zeta_1,\dots,\zeta_K} \max_k \Pr\{S_k \le \zeta_k\} \;\text{ s.t. }\; \Pr\{\cap_{k=1}^K S_k \le \zeta_k\} \ge 1-\alpha. \tag{17}$$

Using (14) and the fact that the CDF is non-decreasing, we can restate (17) as

$$(\widehat{\zeta}_1, \ldots, \widehat{\zeta}_K) = \arg\min_{\zeta_1,\ldots,\zeta_K} \max_k F_{S_k}(\zeta_k) \;\text{ s.t. }\; \Pr\{\cap_{k=1}^K F_{S_k}(S_k) \leq F_{S_k}(\zeta_k)\} \geq 1-\alpha, \tag{18}$$

and further restate it using $\lambda_k \triangleq F_{S_k}(\zeta_k)$ as

$$(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K) = \arg\min_{\lambda_1,\ldots,\lambda_K} \max_k \lambda_k \;\text{ s.t. }\; \Pr\{\cap_{k=1}^K F_{S_k}(S_k) \leq \lambda_k\} \geq 1-\alpha. \tag{19}$$

Although the solution to (19) may not be unique, it suffices to find a single minimax $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K)$. Towards this aim, observe that $\Pr\{\cap_{k=1}^K F_{S_k}(S_k) \leq \lambda_k\}$ is monotonically non-decreasing with respect to any λ_k. Thus, given any tuple $(\lambda_1, \ldots, \lambda_K)$ that satisfies the constraint in (19), the tuple $(\lambda', \ldots, \lambda')$ for $\lambda' \triangleq \max_k \lambda_k$ also satisfies the constraint while simultaneously yielding the same value of the objective "$\max_k \lambda_k$." This implies that we can reframe (19) as a search for a single parameter $\widehat{\lambda}$:

$$\widehat{\lambda} = \arg\min_{\lambda} \lambda \;\text{ s.t. }\; \Pr\{\cap_{k=1}^K F_{S_k}(S_k) \leq \lambda\} \geq 1-\alpha. \tag{20}$$

3.2 Finite-sample case

We now adapt (20) to the practical case where the S_k distributions are unknown and must be learned. For this purpose, we assume access to "tuning" data $\{(u_i, z_i)\}_{i=n+1}^{n+n_{\mathsf{tune}}}$ that is distinct from the data $\{(u_i, z_i)\}_{i=1}^n$ used for split conformal prediction and from the data used to train the predictor $h(\cdot)$ that generates $\widehat{z}_i$.

We propose to do the following for each target component $k \in \{1, \ldots, K\}$. First, we construct the set $d_{\mathsf{tune},k} \triangleq \{(\widehat{z}_{i,k}, z_{i,k})\}_{i=n+1}^{n+n_{\mathsf{tune}}}$ and compute the nonconformity score $s_{i,k}$ for all samples i in $d_{\mathsf{tune},k}$. Using these nonconformity scores, we compute the empirical CDF $\widehat{F}_{S_k}(\cdot)$, where

$$\widehat{F}_{S_k}(\zeta) = \frac{|\{s_{i,k} : s_{i,k} \leq \zeta,\; i = n+1, \ldots, n+n_{\mathsf{tune}}\}|}{n_{\mathsf{tune}}}. \tag{21}$$

Next, we compute the nonconformity scores $s_{i,k}$ for all samples i in the *calibration* set $d_{\mathsf{cal},k} \triangleq \{(\widehat{z}_{i,k}, z_{i,k})\}_{i=1}^n$ and apply the learned $\widehat{F}_{S_k}(\cdot)$ to obtain the transformed calibration scores

$$\overline{s}_{i,k} \triangleq \widehat{F}_{S_k}(s_{i,k}). \tag{22}$$

Because $\overline{s}_{i,k} \in [0,1]$, the scores $\{\overline{s}_{i,k}\}_{i=1}^n$ are normalized across $k \in \{1, \ldots, K\}$. Finally, we take the maximum across components,

$$\overline{s}_i \triangleq \max(\overline{s}_{i,1}, \ldots, \overline{s}_{i,K}), \tag{23}$$

and from these $\{\overline{s}_i\}_{i=1}^n$ compute $\widehat{\lambda}(d_{\mathsf{cal}})$ in the same manner as (2):

$$\widehat{\lambda}(d_{\mathsf{cal}}) = \text{EmpQuant}\left(\tfrac{\lceil (1-\alpha)(n+1) \rceil}{n}; \overline{s}_1, \ldots, \overline{s}_n\right). \tag{24}$$

The target-domain prediction intervals can then be constructed as

$$\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}),k}(\widehat{z}_i) = \left\{z : \widehat{F}_{S_k}\big(s_k(\widehat{z}_{i,k}, z)\big) \leq \widehat{\lambda}(d_{\mathsf{cal}})\right\}, \quad k = 1, \ldots, K. \tag{25}$$

Since the tuning data used to construct $\widehat{F}_{S_k}(\cdot)$ is distinct from the calibration data, our method follows the framework of split conformal prediction. By taking the max of the transformed scores, we ensure the inclusion of all the scores, and thus enjoy the joint marginal coverage guarantee of (9), similar to Sampson & Chan (2024). However, we demonstrate in Section 4 that, as a result of our minimax formulation, our prediction intervals are tighter than those of Sampson & Chan (2024).

Moreover, we can show our method converges to a solution of the minimax optimization in (20) in the limit of infinite tuning and calibration data.

Theorem 1. *For each target component $k = 1, \ldots, K$, suppose that the nonconformity scores $\{S_{i,k}\}_{i=1}^{n+n_{\mathsf{tune}}}$ are i.i.d with CDF $F_{S_k}(\cdot)$, and for $T \triangleq \max_k F_{S_k}(S_k)$, suppose that $F_T(\cdot)$ is continuous and strictly increasing at the $(1-\alpha)$-level quantile of T. Then $\widehat{\lambda}(d_{\mathsf{cal}})$ from (24) converges to λ from (20) almost surely as $n \to \infty$ and $n_{\mathsf{tune}} \to \infty$.*

We provide a proof of Theorem 1 in Appendix A. Through its minimax design, our multi-target approach ensures that no single prediction interval is overly large.

4 Applications of multi-target conformal prediction in imaging

In Section 2.2, we described several applications of conformal prediction to single-target UQ in imaging inverse problems. In this section, we propose several applications of conformal prediction to multi-target UQ in imaging inverse problems.

4.1 Multi-metric blind FRIQ assessment

We first consider blind FRIQ assessment, where the goal is to estimate the FRIQ of a reconstruction $\widehat{x}_0 = f(y_0)$ relative to the true image x_0, given measurements $y_0 = \mathcal{A}(x_0)$ but no direct access to x_0. Whereas Section 2.2 discussed the use of a single FRIQ metric, one may instead be interested in assessing image quality according several FRIQ metrics, since different metrics may be complementary (Wang, 2011).

Consider the case of K FRIQ metrics $\{m_k(\cdot,\cdot)\}_{k=1}^{K}$. To extend the conformal prediction approach from Section 2.2, we set the target vector as $z_i = [z_{i,1}, \ldots, z_{i,K}]^\top \in \mathbb{R}^K$ with $z_{i,k} = m_k(\widehat{x}_i, x_i)$, and use c posterior samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^{c}$ to form the prediction matrix $\widehat{z}_i = [\widehat{z}_{i,1}, \ldots, \widehat{z}_{i,K}]^\top \in \mathbb{R}^{K\times c}$, where $\widehat{z}_{i,k} = [m_k(\widehat{x}_i, \widetilde{x}_i^{(1)}), \ldots, m_k(\widehat{x}_i, \widetilde{x}_i^{(c)})]$. We can also write $\widehat{z}_{i,k} = h_k(u_i)$ for $u_i \triangleq [\widetilde{x}_i^{(1)}, \ldots, \widetilde{x}_i^{(c)}, \widehat{x}_i]^\top$ and an appropriately defined $h_k(\cdot)$. With CQR, the non-conformity score for the kth metric would be computed as

$$s_{i,k} = s_k(\widehat{z}_{i,k}, z_{i,k}) = \max\{\widehat{q}_{\frac{\alpha}{2},k}(\widehat{z}_{i,k}) - z_{i,k}, z_{i,k} - \widehat{q}_{1-\frac{\alpha}{2},k}(\widehat{z}_{i,k})\} \tag{26}$$

with

$$\widehat{q}_{\frac{\alpha}{2},k}(\widehat{z}_{i,k}) = \operatorname{EmpQuant}\left(\frac{\alpha}{2}; \widehat{z}_{i,k}^{(1)}, \ldots, \widehat{z}_{i,k}^{(c)}\right) \quad \text{and} \quad \widehat{q}_{1-\frac{\alpha}{2},k}(\widehat{z}_{i,k}) = \operatorname{EmpQuant}\left(1-\frac{\alpha}{2}; \widehat{z}_{i,k}^{(1)}, \ldots, \widehat{z}_{i,k}^{(c)}\right). \tag{27}$$

With this problem setup, the proposed minimax method from Section 3, or any of the existing multi-target approaches discussed in Section 2.3, can be applied to generate prediction sets $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}),k}(\widehat{z}_0)$ for $k = 1, \ldots, K$. When the joint coverage guarantee in (9) holds, all K prediction sets will simultaneously include the corresponding true-metric values with a probability of at least $1-\alpha$.

4.2 Multi-task uncertainty quantification

Now consider task-based UQ, as described in Section 2.2 for the case of a single task. In practice, one may want to consider several tasks, such as classifying the presence/absence of several different pathologies from a single image. To extend the task-based UQ method from Section 2.2 to K downstream tasks $\{\mu_k(\cdot)\}_{k=1}^{K}$, we form the target vector as $z_i = [z_{i,1}, \ldots, z_{i,K}]^\top \in \mathbb{R}^K$ with $z_{i,k} = \mu_k(x_i)$ and use c posterior samples $\{\widetilde{x}_i^{(j)}\}_{j=1}^{c}$ to form the prediction matrix $\widehat{z}_i = [\widehat{z}_{i,1}, \ldots, \widehat{z}_{i,K}]^\top \in \mathbb{R}^{K\times c}$, where $\widehat{z}_{i,k} = [\mu_k(\widetilde{x}_i^{(1)}), \ldots, \mu_k(\widetilde{x}_i^{(c)})]$. We can also write $\widehat{z}_{i,k} = h_k(u_i)$ for $u_i \triangleq [\widetilde{x}_i^{(1)}, \ldots, \widetilde{x}_i^{(c)}]$ and an appropriately defined $h_k(\cdot)$. With the CQR score (26)-(27), the proposed minimax method from Section 3, or any of the existing multi-target approaches discussed in Section 2.3, can be applied to generate prediction sets $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}),k}(\widehat{z}_0)$ for $k = 1, \ldots, K$. When the joint coverage guarantee (9) holds, all prediction sets will simultaneously contain the corresponding true task-outputs with a probability of at least $1-\alpha$.

4.3 Multi-round measurement acquisition

In Wen et al. (2024) the authors propose a task-based multi-round measurement protocol where measurements are gradually collected until the conformal interval length falls below a user-specified threshold τ. More precisely, at the end of each measurement round $b \in \{1, \dots, B\}$, a prediction $\widehat{z}_0^{[b]} \in \mathbb{R}$ and conformal prediction interval $\mathcal{C}^{[b]}(\widehat{z}_0^{[b]})$ are constructed from the cumulative measurements $y_0^{[b]}$. Measurement collection stops if $|\mathcal{C}^{[b]}(\widehat{z}_0^{[b]})| \leq \tau$ (or if $b = B$) but otherwise continues to the next round. The goal is to reduce measurement costs while guaranteeing that the collected measurements are sufficient for the task. This is especially useful in applications like accelerated MRI, where long scan times increase both patient discomfort and the likelihood of motion artifacts (Knoll et al., 2020).

A limitation of the multi-round protocol from Wen et al. (2024) is that the marginal coverage guarantee holds for each round in isolation, but not for the multi-round protocol as a whole. That is, although $\Pr\{Z_0 \in \mathcal{C}^{[b]}(\widehat{Z}_0^{[b]})\} \geq 1 - \alpha$ for each round b assuming $(Z_0, \widehat{Z}_0^{[b]}), \dots, (Z_n, \widehat{Z}_n^{[b]})$ are exchangeable, we really desire that the multi-round coverage

$$
\begin{aligned}
P_{\mathsf{multi}} &\triangleq \sum_{b=1}^{B} \Pr\left\{Z_0 \in \mathcal{C}^{[b]}(\widehat{Z}_0^{[b]}), \text{final round} = b\right\} && (28)\\
&= \Pr\left\{Z_0 \in \mathcal{C}^{[1]}(\widehat{Z}_0^{[1]}), |\mathcal{C}^{[1]}(\widehat{Z}_0^{[1]})| \leq \tau\right\}\\
&\quad + \sum_{b=2}^{B} \Pr\left\{Z_0 \in \mathcal{C}^{[b]}(\widehat{Z}_0^{[b]}), |\mathcal{C}^{[b]}(\widehat{Z}_0^{[b]})| \leq \tau, |\mathcal{C}^{[b-1]}(\widehat{Z}_0^{[b-1]})| > \tau, \dots |\mathcal{C}^{[1]}(\widehat{Z}_0^{[1]})| > \tau\right\} && (29)
\end{aligned}
$$

is at least $1 - \alpha$. To address this limitation, we note from (28) that

$$
P_{\mathsf{multi}} \geq \sum_{b=1}^{B} \Pr\left\{\, \cap_{k=1}^{B} Z_0 \in \mathcal{C}^{[k]}(\widehat{Z}_0^{[k]}), \text{final round} = b\right\} = \Pr\left\{\, \cap_{k=1}^{B} Z_0 \in \mathcal{C}^{[k]}(\widehat{Z}_0^{[k]})\right\}, \tag{30}
$$

with equality due to the fact that $\sum_{b=1}^{B} \Pr\{\text{final round} = b\} = 1$. Thus, if

$$
\Pr\left\{\, \cap_{b=1}^{B} Z_0 \in \mathcal{C}^{[b]}(\widehat{Z}_0^{[b]})\right\} \geq 1 - \alpha, \tag{31}
$$

then it follows that $P_{\mathsf{multi}} \geq 1 - \alpha$. Since (31) is a special case of the joint marginal coverage guarantee (9), we can ensure (31) using multi-target conformal prediction techniques like the one proposed in Section 3.

That said, the above B-round protocol handles only scalar $z_0 \in \mathbb{R}$, i.e., a single task. To extend it to L tasks $\{\mu_l(\cdot)\}_{l=1}^{L}$, we set the target vector as $z_i = [z_{i,1}, \dots, z_{i,K}]^\top \in \mathbb{R}^K$, where $K = BL$ and $z_{i,L(b-1)+l} = \mu_l(x_i)$ for all b. Then, for each round b, we generate c posterior samples $\{\widetilde{x}_i^{[b](j)}\}_{j=1}^{c}$ via $\widetilde{x}_i^{[b](j)} = g(y_i^{[b]}, \widetilde{v}_i^{[b](j)})$ with i.i.d $\widetilde{v}_i^{[b](j)} \sim \mathcal{N}(0, I)$ and form the predictions $\widehat{z}_i = [\widehat{z}_{i,1}, \dots, \widehat{z}_{i,K}]^\top \in \mathbb{R}^{K \times c}$ with $\widehat{z}_{i,L(b-1)+l} = [\mu_l(\widetilde{x}_i^{[b](1)}), \dots, \mu_l(\widetilde{x}_i^{[b](c)})]$. With the CQR score (26)-(27), the proposed minimax method from Section 3, or any of the existing multi-target approaches discussed in Section 2.3, can be applied. When the joint coverage guarantee in (9) holds, the prediction intervals for all tasks will simultaneously contain their respective targets with probability at least $1 - \alpha$ in the final measurement round.

5 Numerical experiments

We now numerically evaluate the proposed "minimax" multi-target conformal prediction method from Section 3.2, along with the independence-assumption (IA)-based method from Messoudi et al. (2020) and the quantile-normalization (QN)-based method from Sampson & Chan (2024), described in Section 2.3. For $\widehat{q}_{\frac{\alpha}{2},k}(\cdot)$ and $\widehat{q}_{1-\frac{\alpha}{2},k}(\cdot)$, we use the empirical quantile estimator (6) in all cases. For the minimax method, we use the nonconformity score from (7) for each target component. For the IA method, we use the nonconformity score from (7) with an adjusted error-rate of $\alpha_1 = 1 - (1 - \alpha)^{\frac{1}{K}}$ to provide a joint coverage rate of $1 - \alpha$. For the QN method, we compute the nonconformity scores for each target using (12) before taking the max across targets in (10).

We compare all three methods on experiments with accelerated magnetic resonance imaging (MRI) (Knoll et al., 2020; Hammernik et al., 2023). MRI is renowned for its ability to provide high-quality soft tissue images without the use of harmful ionizing radiation. However, MRI scans are slow, which compromises patient comfort and throughput, and can lead to motion artifacts. To mitigate this issue, the scan time is accelerated by a factor of R by collecting only $1/R$ of the measurements required by the Nyquist sampling theorem. Doing so, however, leads to an ill-posed imaging inverse problem, where it is impossible to guarantee recovery of the true image. Thus, for robust diagnoses, uncertainty quantification becomes important.

Data: We follow the experimental setup of Wen et al. (2024), which uses the non-fat-suppressed subset of the multicoil fastMRI knee dataset from Zbontar et al. (2018). This subset contains 17 286 training images and 2188 validation images. To generate the accelerated measurements, the spatial Fourier domain, known as the "k-space", is retrospectively subsampled with random Cartesian masks at acceleration rates $R \in \{16, 8, 4, 2\}$. The masks use Golden Ratio Offset (GRO) sampling (Joshi et al., 2022) and include a fully sampled autocalibration signal (ACS) region in the center, and they are nested such that the measurements collected at each R include all measurements collected at higher R. See Wen et al. (2024) for details.

Models: For the image recovery model $f(\cdot)$, we use the popular E2E-VarNet from Sriram et al. (2020), and for the posterior sampling method $g(\cdot,\cdot)$, we use the conditional normalizing flow (CNF) from Wen et al. (2023) with $c = 32$ posterior samples. Both networks are trained (using the fastMRI training images) to handle acceleration rates $R \in \{16, 8, 4, 2\}$ following the procedure in Wen et al. (2024).

Validation: We first construct a tuning set d_{tune} using 656 of the 2188 fastMRI validation samples (i.e., 30%), selected randomly. Since the joint coverage guarantee (9) holds over the randomness in the calibration and test data, we evaluate performance using $T = 10\,000$ Monte Carlo trials. In each Monte Carlo trial $t \in \{1, \ldots, T\}$, we randomly partition the remaining validation data into a calibration set $d_{\mathsf{cal}}[t]$ of size $n = 1073$ (or 50%) and a test set of size $n_{\mathsf{test}} = 459$ (or 20%) using indices $i \in I_{\mathsf{test}}[t]$. Because the IA and QN methods do not use a tuning set, we merge the tuning samples into their calibration sets (now of size $n + n_{\mathsf{tune}} = 1729$) for fair comparison to the minimax method.

As a coverage metric, we evaluate the empirical joint coverage

$$\mathrm{EJC} \triangleq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_{\mathsf{test}}[t]|}\sum_{i\in\mathcal{I}_{\mathsf{test}}[t]}\prod_{k\in\{1,\ldots,K\}}\mathbb{1}\{z_{i,k}\in\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}[t]),k}(\widehat{z}_i)\}. \tag{32}$$

To quantify the prediction-interval size, we measure the mean interval length for each target k:

$$\mathrm{MIL}_k \triangleq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_{\mathsf{test}}[t]|}\sum_{i\in\mathcal{I}_{\mathsf{test}}[t]}|\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}[t]),k}(\widehat{z}_i)|. \tag{33}$$

5.1 Multi-metric blind FRIQ assessment

We begin with the multi-metric blind FRIQ assessment problem from Section 4.1. For the FRIQ metrics, we consider PSNR, SSIM, learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), and deep image structure and texture similarity (DISTS) (Ding et al., 2020).

Empirical coverage: Table 1 shows EJC versus desired joint coverage $1 - \alpha$ at acceleration $R = 8$. While all methods satisfy the joint coverage guarantee (9), the EJC of the IA method is overly conservative.

Mean interval length: Figure 1 shows mean interval length versus α for each metric at $R = 8$. The minimax method provides the smallest mean interval lengths (i.e., tightest bounds) in nearly every case.

5.2 Multi-task uncertainty quantification

We now consider a multi-task UQ problem, where the goal is to ascertain the presence/absence of each of $K = 5$ different pathologies in accelerated MRI with $R = 8$. We assume that a multi-label soft-output classifier has been trained on clean images x_i to output a vector of probabilities $z_i \in [0, 1]^K$. At inference time, since we have access to only the accelerated measurements y_0 and not the true image x_0, the goal is to construct, for each pathology k, a prediction interval $\mathcal{C}_{\widehat{\lambda}(d_{\mathsf{cal}}),k}(\widehat{z}_0)$ that contains the true soft-output $z_{0,k}$ with some probabilistic guarantee. We apply the minimax, IA, and QN methods as described in Section 4.2.

Table 1: Empirical joint coverage versus $1-\alpha$ for the multi-metric and multi-task MRI experiments at $R=8$.

		$1-\alpha$					
Task	Method	0.70	0.75	0.80	0.85	0.90	0.95
multi-metric	IA (Messoudi et al., 2020)	0.7797	0.8140	0.8485	0.8856	0.9264	0.9625
	QN (Sampson & Chan, 2024)	0.7006	0.7502	0.8006	0.8503	0.9005	0.9503
	Minimax (Ours)	0.7008	0.7511	0.8005	0.8506	0.9002	0.9506
multi-task	IA (Messoudi et al., 2020)	0.7935	0.8344	0.8677	0.8997	0.9340	0.9697
	QN (Sampson & Chan, 2024)	0.7005	0.7502	0.8005	0.8502	0.9005	0.9505
	Minimax (Ours)	0.7013	0.7506	0.8009	0.8511	0.9004	0.9507

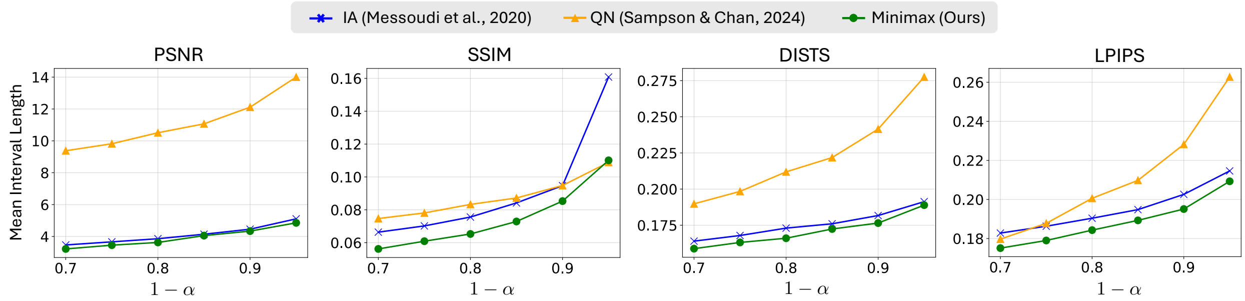


Figure 1: MIL versus $1-\alpha$ for the multi-metric MRI experiments at acceleration $R=8$.

For the classifier, we use a ResNet-50 (He et al., 2016) with $K=5$ outputs in the final linear layer. To train it, we first initialize using ImageNet weights, then pretrain using SimCLR loss (Chen et al., 2020) on the (unlabeled) fastMRI knee data, and finally fine-tuned using binary cross-entropy loss on the (labeled) fastMRI+ knee data (Zhao et al., 2022). The $K=5$ pathologies with the most fastMRI+ samples were chosen for this experiment. Additional details can be found in Appendix B.

Empirical coverage: Table 1 reports EJC versus desired joint coverage $1-\alpha$. As before, all methods satisfy the joint coverage guarantee (9) but the EJC of the IA method is overly conservative. The table shows that the IA method is more conservative in the multi-task experiment than the multi-metric one, which demonstrates how correlation among the nonconformity scores can vary across problems.

Mean interval length: Figure 2 plots MIL versus desired joint coverage $1-\alpha$ for each of the $L=5$ labels. The minimax method produces the smallest average interval lengths (i.e., tightest bounds) except with the meniscus-tear label and the cartilage-thickness-loss label at larger α. Although the QN method gives the best MIL for the meniscus-tear label, it provides the worst MIL for the bone-subchondral-edema label, illustrating a large variability in its performance.

5.3 Multi-round measurement acquisition with FRIQ guarantees

We now consider applying the multi-round measurement protocol from Section 4.3 to accelerate MRI while providing a probabilistic FRIQ guarantee. We adopt the experimental setup of Wen et al. (2025), where measurements are collected over $B=5$ rounds at acceleration rates $R \in \{16,8,4,2,1\}$ but stop as soon as a conformal upper bound on DISTS[1] falls below a threshold of $\tau = 0.16$. To adapt the proposed multi-target method from Section 4.3 to this upper-bounding setup, we run the IA and minimax methods with the nonconformity score $s_k(\widehat{z}_{i,k}, z_{i,k}) = z_{i,k} - \widehat{q}_{1-\frac{\alpha}{2}}(\widehat{z}_{i,k})$ and run QN with $s_k(\widehat{z}_{i,k}, z_{i,k}) = \big(z_{i,k} - \widehat{q}_{1-\frac{\alpha}{2},k}(\widehat{z}_i)\big)\frac{\widehat{q}_{1-\frac{\alpha}{2},1}(\widehat{z}_i)-\widehat{q}_{\frac{\alpha}{2},1}(\widehat{z}_i)}{\widehat{q}_{1-\frac{\alpha}{2},k}(\widehat{z}_i)-\widehat{q}_{\frac{\alpha}{2},k}(\widehat{z}_i)}$, as motivated by (12).

[1] A recent clinical MRI study (Kastryulin et al., 2023) evaluated 35 FRIQ metrics and found that DISTS correlated best with radiologists' ratings of perceived noise level, contrast level, and artifacts when comparing reconstructions to ground-truth images.

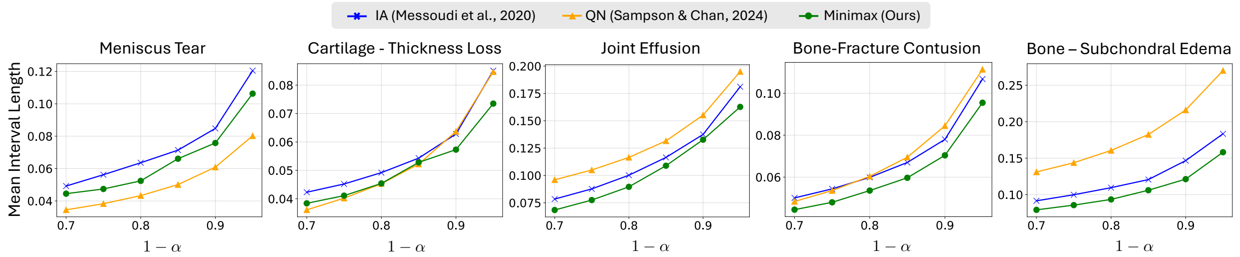
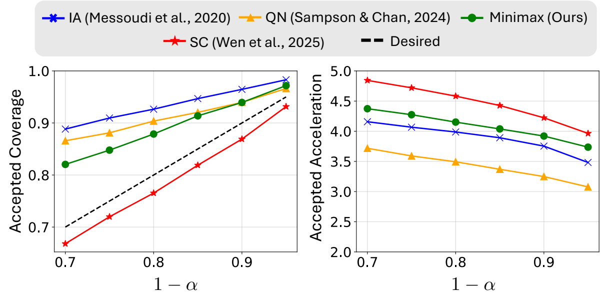


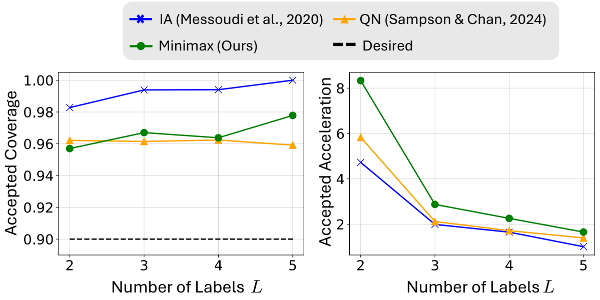
Figure 2: MIL versus $1 - \alpha$ for multi-label classification with MRI acceleration $R = 8$.



(a) Empirical accepted coverage

(b) Average accepted acceleration

Figure 3: For multi-round measurements that stop as soon as the DISTS upper bound falls below $\tau = 0.16$, (a) plots EAC and (b) plots R_{avg} versus the desired coverage $1 - \alpha$.



(a) Empirical accepted coverage

(b) Average accepted acceleration

Figure 4: For multi-round measurements that stop as soon as the prediction intervals for all pathology labels fall below $\tau = 0.1$, (a) plots EAC and (b) plots R_{avg} versus the number of labels L at $1 - \alpha = 0.9$.

Using b_i to denote the final round for test sample i, Fig. 3(a) plots the empirical accepted coverage

$$\text{EAC} \triangleq \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\mathcal{I}_{\text{test}}[t]|} \sum_{i \in \mathcal{I}_{\text{test}}[t]} \mathbb{1}\{z_i \in \mathcal{C}^{[b_i]}(\widehat{z}_i^{[b_i]})\}, \tag{34}$$

versus the desired coverage $1 - \alpha$ for the IA, QN, and minimax versions of the multi-target method from Section 4.3, as well as the separate calibration (SC) method from Wen et al. (2025). The figure shows that the measurements accepted by the IA-, QN-, and minimax-based multi-target EACs provide the desired coverage, while those accepted by the SC method do not. Figure 3(b) plots the average accepted acceleration-rate

$$R_{\text{avg}} \triangleq \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\mathcal{I}_{\text{test}}[t]|} \sum_{i \in \mathcal{I}_{\text{test}}[t]} \frac{1}{R_{b_i}} \right)^{-1} \tag{35}$$

versus $1 - \alpha$, where R_{b_i} is the acceleration rate of the final round for test sample i. Although the SC method achieves higher R_{avg} than the IA and minimax methods, it comes at the cost of not providing a coverage guarantee. Among the multi-target methods, the minimax method yields the highest R_{avg} in all cases, demonstrating the advantage of tight conformal bounds.

5.4 Multi-round measurement acquisition with downstream classification guarantees

Finally, we consider an application of the multi-round measurement protocol from Section 4.3 that aims to accelerate MRI while providing a probabilistic guarantee on downstream classification of multiple pathologies. For this we combine the multi-round setup from Section 5.3 with the multi-label setup from Section 5.2. In particular, we take measurements over $B = 5$ rounds at acceleration rates $R \in \{16, 8, 4, 2, 1\}$ but stop as soon as the prediction interval lengths for all L pathology labels fall below $\tau = 0.1$. Rather that exclusively considering $L = 5$ pathology labels, we experiment with $L \in \{2, 3, 4, 5\}$ to investigate the effect of L. Here, $L = l$ corresponds to the l most prevalent classes in the fastMRI+ training data.

For a desired coverage of $1 - \alpha = 0.9$, Fig. 4(a) plots EAC versus the number of labels L. The figure again shows that the IA, QN, and minimax methods meet the desired coverage in all cases. Figure 4(b) plots R_{avg} versus the number of labels L. The figure shows that the proposed minimax method yields the highest R_{avg} across all values of L. It also shows that R_{avg} decreases with L, which is expected because the stopping criterion becomes more strict with larger L.

6 Conclusion

Motivated by the need for multi-target uncertainty quantification in imaging inverse problems, we propose a minimax approach to multi-target conformal prediction. The proposed method aims to minimize the maximum single-target coverage, $\max_k \Pr\{Z_{0,k} \in \mathcal{C}_{\widehat{\lambda}(D_{\text{cal}}),k}(\widehat{Z}_0)\}$, across targets k subject to a joint coverage guarantee of the form $\Pr\{\cap_{k=1}^{K} Z_{0,k} \in \mathcal{C}_{\widehat{\lambda}(D_{\text{cal}}),k}(\widehat{Z}_0)\} \geq 1 - \alpha$, where α is user-specified. We proved that the finite-sample version of our approach converges to the desired minimax solution as the tuning and calibration sets grow in size. Furthermore, for inverse problems, we proposed a multi-round measurement acquisition scheme with marginal coverage guarantees on the final-round prediction intervals. We numerically compared the proposed method to several existing multi-target conformal prediction methods on four accelerated-MRI problems and found that the proposed minimax method gives tighter prediction intervals in most cases while guaranteeing joint marginal coverage.

Limitations

There are several limitations to this work. First, like with many conformal prediction methods, the joint-coverage guarantee (9) is known to hold only for statistically exchangeable prediction/target pairs $\{(\widehat{Z}_i, Z_i)\}_{i=0}^{n}$. Furthermore, the convergence of the finite-sample method has been established only under i.i.d. nonconformity scores $\{S_{i,k}\}_{i=1}^{n+n_{\text{tune}}}$. Further work is needed to generalize these restrictions, and the works Tibshirani et al. (2019), Barber et al. (2023), Cauchois et al. (2024) suggest modifications that address non-exchangeability. In addition, the proposed applications to MRI are preliminary in that rigorous clinical trials are needed before they are adopted in practice.

Broader impact statement

We expect that our methodology will positively impact the field of imaging inverse problems by providing prediction intervals on multiple estimation targets that involve the (unknown) true image. These intervals inform the practitioner of how much uncertainty the measurement-and-reconstruction process introduces to downstream tasks, and whether the collected measurements are sufficient for a given reconstruction method. Furthermore, the propose multi-round acquisition protocol allows one to collect fewer measurements while still providing a guarantee on estimation performance.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Amit P. Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proc. Intl. Conf. on Machine Learning*, 2022. doi: 10.48550/arXiv.2202.05265.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, June 2019.

Christopher R. S. Banerji, Tapabrata Chakraborti, Chris Harbron, and Ben D. MacArthur. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, 29(12):2996–2998, 2023. doi: 10.1038/s41591-023-02562-7.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 46:3321–3333, 2023.

Chinmay Belthangady and Loic A Royer. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nature Methods*, 16(12):1215–1225, 2019.

Mario Bertero, Patrizia Boccacci, and Christine De Mol. *Introduction to Inverse Problems in Imaging.* CRC press, 2021.

Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Trans. on Medical Imaging*, 40(11):3249–3260, 2021.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pp. 1–66, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Intl. Conf. on Machine Learning*, pp. 1597–1607, 2020.

Matt Y Cheung, Tucker J Netherton, Laurence E Court, Ashok Veeraraghavan, and Guha Balakrishnan. Metric-guided image reconstruction bounds via conformal prediction. *arXiv:2404.15274*, 2024.

Linda C Chu, Anima Anandkumar, Hoo Chang Shin, and Elliot K Fishman. The potential dangers of artificial intelligence for radiology and radiologists. *Journal of the American College of Radiology*, 17(10): 1309–1311, 2020.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Proc. Intl. Conf. on Medical Image Computation and Computer-Assisted Intervention*, pp. 529–536, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Bert E Fristedt and Lawrence F Gray. *A Modern Approach to Probability Theory.* Springer, 2013.

Nina M Gottschling, Vegard Antun, Anders C Hansen, and Ben Adcock. The troublesome kernel—On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems. *arXiv:2001.01258*, 2023.

Kerstin Hammernik, Thomas Küstner, Burhaneddin Yaman, Zhengnan Huang, Daniel Rueckert, Florian Knoll, and Mehmet Akçakaya. Physics-driven deep learning for computational magnetic resonance imaging: Combining physics and machine learning for improved medical imaging. *IEEE Signal Processing Magazine*, 40(1):98–114, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

David P Hoffman, Isaac Slavitt, and Casey A Fitzpatrick. The promise and peril of deep learning in microscopy. *Nature Methods*, 18(2):131–132, 2021.

Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv.2211.09795*, 2022. doi: 10.48550/arXiv.2211.09795.

Mihir Joshi, Aaron Pruitt, Chong Chen, Yingmin Liu, and Rizwan Ahmad. Technical report (v1.0)–pseudo-random cartesian sampling for dynamic MRI. *arXiv:2206.03630*, 2022.

Sergey Kastryulin, Jamil Zakirov, Nicola Pezzotti, and Dmitry V Dylov. Image quality assessment for magnetic resonance imaging. *IEEE Access*, 11:14154–14168, 2023.

Florian Knoll, Kerstin Hammernik, Chi Zhang, Steen Moeller, Thomas Pock, Daniel K Sodickson, and Mehmet Akcakaya. Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues. *IEEE Signal Processing Magazine*, 37(1):128–140, January 2020.

Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *Proc. Intl. Conf. on Learning Representations*, 2023.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society*, 76, 2014. doi: 10.1111/rssb.12021.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2018.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Conformal multi-target regression using neural networks. In *Proc. Symp. on Conformal and Probabilistic Prediction With Applications*, pp. 65–83. PMLR, 2020.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Proc. Symp. on Conformal and Probabilistic Prediction With Applications*, pp. 294–306. PMLR, 2022.

Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. *IEEE Trans. on Medical Imaging*, 40(9):2306–2317, 2021.

Dominik Narnhofer, Andreas Habring, Martin Holler, and Thomas Pock. Posterior-variance-based error quantification for inverse problems in imaging. *SIAM Journal on Imaging Sciences*, 17(1):301–333, 2024.

Roger Nelsen. *An Introduction to Copulas.* Springer NY, 2nd edition, 2006.

Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, May 2020.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Proc. European Conf. on Machine Learning*, pp. 345–356, 2002. doi: 10.1007/3-540-36755-1_29.

Ji Won Park, Robert Tibshirani, and Kyunghyun Cho. Semiparametric conformal prediction. *arXiv:2411.02114*, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Proc. Neural Information Processing Systems Conf.*, pp. 3543–3553, 2019. doi: 10.48550/arXiv.1905.03222.

Max Sampson and Kung-Sik Chan. Conformal multi-target hyperrectangles. *Statical Analysis and Data Mining*, 17(5), 2024.

Swami Sankaranarayanan, Anastasios N. Angelopoulos, Stephen Bates, Yaniv Romano, and Phillip Isola. Semantic uncertainty intervals for disentangled latent spaces. In *Proc. Neural Information Processing Systems Conf.*, 2022. doi: 10.48550/arXiv.2207.10074.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *Proc. Intl. Conf. on Medical Image Computation and Computer-Assisted Intervention*, pp. 64–73, 2020.

Jacopo Teneggi, Matthew Tivnan, J. Webster Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control, 2023.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Proc. Neural Information Processing Systems Conf.*, volume 32, 2019.

Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination index: An image quality metric for generative reconstruction models. In *Proc. Intl. Conf. on Medical Image Computation and Computer-Assisted Intervention*, pp. 449–458, 2024.

Sumeet Virmani, Rashmi Virmani, Jagadeesh Singh, and Amjad Ali. Imaging artifacts: A pictorial review. *Journal of Nuclear Medicine*, 56, May 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World.* Springer, New York, 2005.

Zhou Wang. Applications of objective image quality assessment methods. *IEEE Signal Processing Magazine*, 28(6):137–142, 2011.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, April 2004.

Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. A conditional normalizing flow for accelerated multi-coil MR imaging. In *Proc. Intl. Conf. on Machine Learning*, 2023.

Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. Task-driven uncertainty quantification in inverse problems via conformal prediction. In *Proc. European Conf. on Computer Vision*, 2024.

Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. Conformal bounds on full-reference image quality for imaging inverse problems. *Trans. on Machine Learning Research*, 2025. ISSN 2835-8856.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv:1811.08839*, 2018.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W. Lui, Michael S. Hansen, and Matthew P. Lungren. fastMRI+: Clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 9(1):152, 2022. doi: 10.1038/s41597-022-01255-z.

A Proof of Theorem 1

In this section, we show that $\widehat{\lambda}(d_{\mathsf{cal}})$ in (24) converges to $\widehat{\lambda}$ in (20) as $n \to \infty$ and $n_{\mathsf{tune}} \to \infty$. We first recall the definition of almost-sure convergence.

Definition 1 (Almost-sure convergence)**.** *Let $(X_n)_{n \geq 1}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. We say that X_n* ***converges almost surely*** *(or with probability 1) to a random variable X, denoted as $X_n \xrightarrow{a.s.} X$, if*

$$\Pr\left\{ \lim_{n \to \infty} X_n = X \right\} = 1.$$

That is, the outcomes $\omega \in \Omega$ under which $X_n(\omega)$ converges to $X(\omega)$ occur with probability one.

We now state two theorems that form the basis for our convergence analysis.

Theorem 2 (Glivenko-Cantelli (Fristedt & Gray, 2013))**.** *Suppose $X_1, \ldots, X_n$ are i.i.d random variables with CDF $F(\cdot)$. Define the empirical CDF as*

$$\widehat{F}(x) \triangleq \frac{|\{i : X_i \leq x,\ i = 1, \ldots, n\}|}{n}.$$

Then $\widehat{F}(\cdot)$ converges uniformly to $F(\cdot)$ almost surely, i.e.

$$\sup_{x \in \mathbb{R}} |\widehat{F}(x) - F(x)| \xrightarrow{a.s.} 0.$$

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables with CDF $F(\cdot)$. Define the quantile at level $p \in (0, 1)$ as*

$$Q(p) = \inf\{x : F(x) \geq p\}$$

and the empirical quantile at level $p \in (0, 1)$ as

$$Q_n(p) = \inf\{x : \widehat{F}(x) \geq p\},$$

where $\widehat{F}(\cdot)$ is the empirical CDF. For a fixed level p, construct the n-dependent level

$$\gamma_n = \frac{\lceil p(n+1) \rceil}{n},$$

which approaches p as $n \to \infty$. If $F(\cdot)$ is continuous and strictly increasing at $Q(p)$, then

$$Q_n(\gamma_n) \xrightarrow{a.s} Q(p).$$

That is, the empirical quantile at level γ_n converges almost surely to the true quantile at level p.

Proof. First, we analyze the convergence of γ_n. Observe that

$$\gamma_n = \frac{\lceil p(n+1) \rceil}{n} = \frac{p(n+1) + \Delta_n}{n} = p + \frac{p + \Delta_n}{n} \tag{36}$$

where $\Delta_n \in [0, 1)$ accounts for the rounding of the ceiling function. Thus

$$p < \gamma_n < p + \frac{2}{n}, \tag{37}$$

and $\lim_{n \to \infty} \gamma_n = p$. Next, we look to bound $Q_n(\gamma_n)$ as $n \to \infty$. For any fixed $\epsilon > 0$, and assuming $F(\cdot)$ is continuous and strictly increasing at $Q(p)$, we have

$$F(Q(p) - \epsilon) < p < F(Q(p) + \epsilon).$$

And by Theorem 2, the Glivenko-Cantelli theorem, $\widehat{F}(\cdot)$ converges uniformly to $F(\cdot)$ almost surely. This means that, for any $\delta > 0$, there almost surely exists an N such that, for all $n \geq N$ and for all $x \in \mathbb{R}$,

$$|\widehat{F}(x) - F(x)| \leq \delta.$$

By choosing

$$\delta < \min\{p - F(Q(p) - \epsilon), F(Q(p) + \epsilon) - p\}$$

we get

$$F(Q(p) - \epsilon) + \delta < p < F(Q(p) + \epsilon) - \delta, \tag{38}$$

and so

$$\widehat{F}(Q(p) - \epsilon) < p < \widehat{F}(Q(p) + \epsilon) \tag{39}$$

for all $n \geq N$. We now establish two intermediate results.

Lemma 4. *For sufficiently large n, we have $Q_n(\gamma_n) \geq Q(p) - \epsilon$.*

Proof. We prove the claim using contradiction. Suppose that $Q_n(\gamma_n) < Q(p) - \epsilon$. Then, due to the non-decreasing property of $\widehat{F}(\cdot)$, we have

$$\widehat{F}(Q_n(\gamma_n)) \leq \widehat{F}(Q(p) - \epsilon) \tag{40}$$

for any n. Furthermore, since $\widehat{F}(Q_n(\gamma_n)) \geq \gamma_n$ for any n by the definition of the empirical quantile, and since $\gamma_n > p$ from (37), we have

$$\widehat{F}(Q(p) - \epsilon) \geq \gamma_n > p. \tag{41}$$

However, (41) contradicts (39) when $n \geq N$. This implies that $Q_n(\gamma_n) \geq Q(p) - \epsilon$ for sufficiently large n. ∎

Lemma 5. *For sufficiently large n, we have $Q_n(\gamma_n) \leq Q(p) + \epsilon$.*

Proof. We prove the claim using contradiction. Suppose that $Q_n(\gamma_n) > Q(p) + \epsilon$. Recall that, by definition, $Q_n(\gamma_n) = \inf\{x : \widehat{F}(x) \geq \gamma_n\}$. Thus if $Q(p) + \epsilon < Q_n(\gamma_n)$ then

$$\widehat{F}(Q(p) + \epsilon) < \gamma_n. \tag{42}$$

And recall from (38) that $F(Q(p) + \epsilon) - \delta > p$, or equivalently that

$$F(Q(p) + \epsilon) - \frac{\delta}{2} > p + \frac{\delta}{2}. \tag{43}$$

From Theorem 2, the Glivenko-Cantelli theorem, $\widehat{F}(\cdot)$ converges uniformly to $F(\cdot)$ almost surely. This means that, for the given δ, there almost surely exists an N' such that, for all $n \geq N'$ and any $x \in \mathbb{R}$,

$$|\widehat{F}(x) - F(x)| \leq \frac{\delta}{2} \quad \Rightarrow \quad \widehat{F}(x) \geq F(x) - \frac{\delta}{2}. \tag{44}$$

Combining (43) and (44), we have that, for all $n \geq N'$,

$$\widehat{F}(Q(p) + \epsilon) \geq F(Q(p) + \epsilon) - \frac{\delta}{2} > p + \frac{\delta}{2}.$$

From (37), we see that, for $n \geq 4/\delta$,

$$\gamma_n < p + \frac{2}{n} \leq p + \frac{\delta}{2}.$$

Thus, for sufficiently large n, we have

$$\widehat{F}(Q(p) + \epsilon) > \gamma_n,$$

which contradicts (42). This implies that $Q_n(\gamma_n) \leq Q(p) + \epsilon$ for large n. ∎

Lemma 4 and Lemma 5 hold almost surely for an arbitrary $\epsilon > 0$, and together say that

$$Q(p) - \epsilon \leq Q_n(\gamma_n) \leq Q(p) + \epsilon$$

for sufficiently large n. Since we can make ϵ arbitrarily small, we have that

$$\lim_{n\to\infty} Q_n(\gamma_n) = Q(p),$$

almost surely, and thus $Q_n(\gamma_n) \xrightarrow{a.s.} Q(p)$. □

Having established Theorem 2 and Theorem 3, we now return to our main objective, which is proving that the $\widehat{\lambda}(d_{\mathsf{cal}})$ in (24) converges to the $\widehat{\lambda}$ in (20). For clarity, we restate Theorem 1 here.

Theorem (Restatement of Theorem 1)**.** *For each target component $k = 1, \dots, K$, suppose that the nonconformity scores $\{S_{i,k}\}_{i=1}^{n+n_{\mathsf{tune}}}$ are i.i.d with CDF $F_{S_k}(\cdot)$, and for $T \triangleq \max_k F_{S_k}(S_k)$, suppose that $F_T(\cdot)$ is continuous and strictly increasing at the $(1-\alpha)$-level quantile of T. Then $\widehat{\lambda}(d_{\mathsf{cal}})$ from (24) converges to $\widehat{\lambda}$ from (20) almost surely as $n \to \infty$ and $n_{\mathsf{tune}} \to \infty$.*

Proof. We first analyze the effect of $n_{\mathsf{tune}} \to \infty$ for an arbitrary fixed n. Recall that the empirical CDF $\widehat{F}_{S_k}(\cdot)$ of the nonconformity score for the kth component is computed as in (21) using the tuning samples $\{S_{i,k}\}_{i=n+1}^{n+n_{\mathsf{tune}}}$. From Theorem 2, $\widehat{F}_{S_k}(\cdot)$ converges uniformly to the CDF $F_{S_k}(\cdot)$ almost surely as $n_{\mathsf{tune}} \to \infty$. As a result, it follows that for each *calibration* nonconformity score $S_{i,k}$, where $i \in \{1, \dots, n\}$, we have

$$\overline{S}_{i,k} \triangleq \widehat{F}_{S_k}(S_{i,k}) \xrightarrow{a.s.} F_{S_k}(S_{i,k})$$

as $n_{\mathsf{tune}} \to \infty$, recalling the definition of the transformed score $\overline{S}_{i,k}$ from (22). Let us now consider the maximum transformed score $\overline{S}_i \triangleq \max_k \{\overline{S}_{i,k}\}_{k=1}^K$ defined in (23). Since the maximum function is continuous everywhere on $\mathbb{R}^K$ and $\widehat{F}_{S_k}(S_{i,k}) \xrightarrow{a.s.} F_{S_k}(S_{i,k})$, the continuous mapping theorem implies that

$$\overline{S}_i = \max_k \widehat{F}_{S_k}(S_{i,k}) \xrightarrow{a.s.} \max_k F_{S_k}(S_{i,k}) \triangleq T_i$$

as $n_{\mathsf{tune}} \to \infty$. Because $\{S_{i,k}\}_{i=1}^n$ are assumed to be i.i.d with CDF $F_{S_k}(\cdot)$, we see that $\{T_i\}_{i=1}^n$ are i.i.d with CDF $F_T(\cdot)$ for $T \triangleq \max_k F_{S_k}(S_k)$.

Next, we analyze the effect of $n \to \infty$. Let us denote the n-sample empirical quantile of T as $Q_n(\cdot)$ and the quantile of T as $Q(\cdot)$. Recall from (24) that

$$\widehat{\lambda}(D_{\mathsf{cal}}) = Q_n\bigg(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}\bigg).$$

Because $F_T(\cdot)$ is assumed to be continuous and strictly increasing at $Q(1-\alpha)$, Theorem 3 establishes that, as $n \to \infty$,

$$\widehat{\lambda}(D_{\mathsf{cal}}) = Q_n\bigg(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}\bigg) \xrightarrow{a.s.} Q(1-\alpha). \tag{45}$$

Finally, recall the definition of $\widehat{\lambda}$ from (20):

$$\widehat{\lambda} = \arg\min_{\lambda} \lambda \;\text{ s.t. }\; \Pr\{\cap_{k=1}^K F_{S_k}(S_k) \leq \lambda\} \geq 1-\alpha.$$

The constraint can be rewritten as

$$\Pr\{\max_k F_{S_k}(S_k) \leq \lambda\} = \Pr\{T \leq \lambda\} \geq 1-\alpha,$$

which allows (20) to be rewritten as

$$\widehat{\lambda} = \arg\min_{\lambda} \lambda \;\text{ s.t. }\; \Pr\{T \leq \lambda\} \geq 1-\alpha. \tag{46}$$

Table 2: Number of positive samples in the non-fat-suppressed subset of the fastMRI+ knee dataset.

Label	Positive Training Samples	Positive Validation Samples
Meniscus Tear	1921	335
Cartilage - Partial Thickness loss/defect	871	176
Joint Effusion	225	41
Bone-Fracture/Contusion/dislocation	97	6
Bone - Subchondral edema	76	21

Table 3: Classifier performance on the fastMRI+ validation set.

Label	Accuracy	Precision	Recall	AUROC
Meniscus Tear	0.6595	0.3005	0.9784	0.889
Cartilage - Partial Thickness loss/defect	0.6184	0.1558	0.8988	0.8564
Joint Effusion	0.9031	0.1356	0.8000	0.9465
Bone-Fracture/Contusion/dislocation	0.7715	0.0060	0.5000	0.7971
Bone - Subchondral edema	0.5704	0.0127	0.5714	0.6338
Average	0.7046	0.1221	0.7497	0.8246

But the $\widehat{\lambda}$ in (46) is simply the $(1-\alpha)$-level quantile of T. In other words,

$$\widehat{\lambda} = \inf\{\lambda : F_T(\lambda) \geq 1-\alpha\} = Q(1-\alpha). \tag{47}$$

Finally, combining (45) with (47), we conclude that

$$\widehat{\lambda}(D_{\mathsf{cal}}) \xrightarrow{a.s.} \widehat{\lambda}$$

as $n_{\mathsf{tune}} \to \infty$ and $n \to \infty$. □

B Classifier Details

We train the multi-label classifier on the $K = 5$ labels with the most annotations in the non-fat-suppressed subset of the fastMRI+ knee data from Zhao et al. (2022). Table 2 shows the number of positive samples for each of those labels. Note that images with multiple instances of the same pathology only count as a single positive sample.

We implement and train the multi-label classifier using nearly the same procedure as Wen et al. (2024). In particular, we start by initializing a standard ResNet-50 (He et al., 2016) with the pretrained ImageNet weights from (Deng et al., 2009), after which we reduce the number of final-layer outputs to $K = 5$. Then we pretrain the network in a self-supervised fashion using the (unlabeled) non-fat-suppressed fastMRI knee data following the SimCLR procedure from Chen et al. (2020) with a learning rate of 0.0002, batch size of 128, and 500 epochs. Finally, we perform supervised fine-tuning using binary cross-entropy loss on the fastMRI+ data, where we address class imbalance by weighting the loss contribution from each class by the ratio of negative labels to positive labels for that particular class. To encourage adversarial robustness, we use the same l_2-bounded gradient ascent attack as Wen et al. (2024), and we train the classifier for 150 epochs with a batch size of 128, learning rate of 5e−5, and weight decay of 1e−7. Finally, we save the model checkpoint with the lowest validation loss. Performance on the validation dataset is shown in Table 3.