Metric-Guided Conformal Bounds for Probabilistic Image Reconstruction

Matt Y. Cheung[1,2][0000−0002−7846−3297], Tucker J. Netherton[2][0000−0003−1583−7121], Laurence E. Court[2][0000−0002−3241−6145], Ashok Veeraraghavan[1][0000−0001−5043−7460], and Guha Balakrishnan[1][0000−0001−8703−1368]

[1] Department of Electrical & Computer Engineering, Rice University, Houston TX, USA

[2] Department of Radiation Physics, The University of Texas M.D. Anderson Cancer Center, Houston TX, USA

guha@rice.edu

Abstract. Modern deep learning reconstruction algorithms generate impressively realistic scans from sparse inputs, but can often produce significant inaccuracies. This makes it difficult to provide statistically guaranteed claims about the true state of a subject from scans reconstructed by these algorithms. In this study, we propose a framework for computing provably valid prediction bounds on claims derived from probabilistic black-box image reconstruction algorithms. The key insights are to represent reconstructed scans with a derived clinical metric of interest and to calibrate bounds on the ground truth metric with conformal prediction (CP) using a prior calibration dataset. These bounds convey interpretable feedback about the subject's state, and can also be used to retrieve nearest-neighbor reconstructed scans for visual inspection. We demonstrate the utility of this framework on sparse-view computed tomography (CT) for fat mass quantification and radiotherapy planning tasks. Results show that our framework produces visual bounds with better semantical interpretation than conventional pixel-based bounding approaches and captures important spatial correlations. Furthermore, we can flag dangerous outlier reconstructions that look plausible but have statistically unlikely metric values. Code available at: https://github.com/matthewyccheung/conformal-metric

Keywords: Image Reconstruction · Conformal Prediction · Sparse-view CT · Deep Generative Models

1 Introduction

Classical sparse medical imaging reconstruction algorithms – from Iterative Reconstruction (IR) methods [15] used in computed tomography (CT) to compressive sensing methods [22] used in magnetic resonance imaging (MRI) – have one extremely desirable property: when they fail, they ***fail***. That is, in most cases,

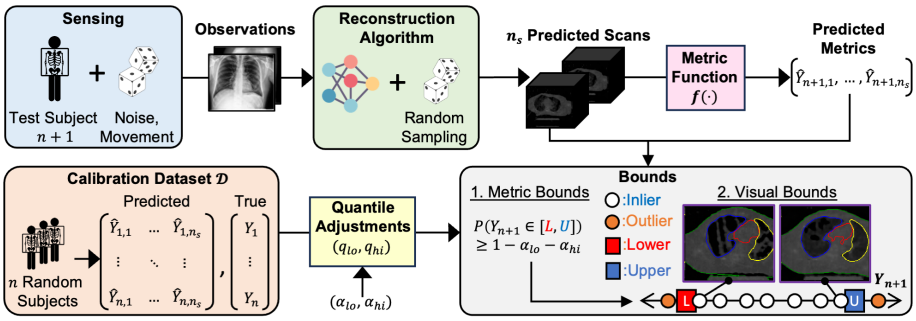

Fig. 1. Overview of our approach. Standard imaging pipeline with probabilistic image acquisition, probabilistic image reconstruction algorithm, reconstructed volumes, and a downstream application. We propose to use CP on downstream metrics to attain meaningful lower and upper bounds, inliers, and outliers. Lower and upper bound images correspond to the reconstructions with metrics closest to the CP prediction interval lower and upper bounds (nearest neighbors or NN). CP prediction intervals are constructed based on a separate calibration set.

"incorrect" reconstructions also look obviously poor to the naked eye, and we understand their modes of failure. In contrast, modern deep generative and neural rendering algorithms [17, 20, 34] use rich implicit or data-driven priors to produce impressive looking images, but are known to *hallucinate*, i.e., make predictions which are inaccurate but look plausible [4, 12, 23]. This creates an operational challenge in engendering trust for these algorithms in medical imaging.

In the typical image reconstruction pipeline, variations in observation acquisition (e.g., patient movement, sensor noise) [9, 18] and lossiness of observations contribute to reconstruction uncertainty. Deep learning-based reconstruction algorithms can account for some of this uncertainty by randomizing inputs or parameters to produce a distribution of solutions, but these approaches are not guaranteed to capture the full space of solutions. In particular, *epistemic uncertainty* caused by missing data in the training distribution is likely to be underestimated and contribute to hallucinations [5].

This raises a crucial question: For a given test subject, given a random sample of reconstructed scans produced by a black-box deep learning algorithm, is it possible to make statistically guaranteed claims about the subject's true state? We show that this is possible and propose a framework based on two ideas (Fig. 1). First, we make such statistical claims over a meaningful downstream metric derived from the scans, e.g., heart volume or fat content, rather than over raw pixel values, which suffer from high dimensionality and limited interpretability. Second, we assume an available *calibration dataset* consisting of metric values derived from prior reconstructed and ground-truth scans. At test time, this calibration dataset may be used to adjust for systematic deviations of metrics derived from the reconstructions with respect to ground-truth scans.

Algorithm 1 Metric-Guided Bound Computation.

Inputs: Calibration set $\mathcal{D} = \{\hat{Y}_i\}_{i=1}^n$, test subject sample metrics $\hat{Y}_{n+1}$, lower and upper miscoverage rates α_{lo} and α_{hi}.
Outputs: Prediction interval $C(\hat{Y}_{n+1})$, lower and upper bound reconstructions $L(\hat{I}_{n+1})$ and $U(\hat{I}_{n+1})$, inlier and outlier reconstructions $In(\hat{I}_{n+1})$ and $Out(\hat{I}_{n+1})$.

▷ Perform calibration to get quantile adjustments
for $i = 1 : n$ **do**
  $[s_{i,lo}, s_{i,hi}] \leftarrow [Q_{\alpha_{lo}}(\hat{Y}_i) - Y_i, Y_i - Q_{1-\alpha_{hi}}(\hat{Y}_i)]$
end for
$\hat{\alpha}_{lo}, \hat{\alpha}_{hi} \leftarrow \lfloor \alpha_{lo}(n+1) \rfloor / n, \lfloor \alpha_{hi}(n+1) \rfloor / n$
$[q_{lo}, q_{hi}] \leftarrow [Q_{1-\hat{\alpha}_{lo}}(\{s_{i,lo}\}_{i=1}^n), Q_{1-\hat{\alpha}_{hi}}(\{s_{i,hi}\}_{i=1}^n)]$

▷ Compute prediction interval for test patient.
$C(\hat{Y}_{n+1}) \leftarrow [Q_{\alpha_{lo}}(\hat{Y}_{n+1}) - q_{lo}, Q_{1-\alpha_{hi}}(\hat{Y}_{n+1}) + q_{hi}]$

More specifically, we first algorithmically compute a metric of interest from each sample in a set of n_s reconstructed scans for a test subject (Fig. 1-top). Second, using this set of metric values, we compute valid prediction intervals that reliably contain the ground truth value on average using a calibration dataset $\mathcal{D}$ of n random subjects (Fig. 1-bottom). To do so, we build on conformal prediction (CP), a powerful technique to construct uncertainty-aware prediction intervals for values predicted by a black-box algorithm [11, 24, 28, 30]. Under the assumption of exchangeability, the resulting bounds will provably contain the ground truth values $(1 - \alpha_{lo} - \alpha_{hi})\%$ of the time for some set lower/upper mis-coverage rates α_{lo}/α_{hi}. Finally, we can link the lower/upper metric bounds to their nearest reconstructed scans in the sample set to provide visual interpretation.

Most existing strategies applying CP to image reconstruction operate directly on pixel values [2, 16, 29], which yield bounds with limited interpretability and do not capture spatial correlations. Other studies propose using low-dimensional representations of pixels computed using principal component analysis (PCA) [3] or latent spaces of pretrained generative models [27]. However, PCA is expensive to compute for scans larger than a small size, and high-quality generative models are not available for arbitrary medical image sets. In contrast, we propose computing CP bounds over metrics derived from the reconstructed scans, resulting in a flexible, interpretable, and computationally tractable solution.

We demonstrate the utility of our method on sparse-view CT reconstruction. We focus specifically on the downstream metrics for fat mass quantification and radiotherapy (RT) planning using de-identified CT datasets from The University of Texas M.D. Anderson Cancer Center. Our results first show that our method achieves valid coverage for downstream metrics while more common pixel-based bounding methods exhibit significant undercoverage, i.e., generated lower and upper bounds do not contain the ground truth metric with a user-specified probability. Second, we demonstrate that meaningful visual bounds in the form of nearest neighbor reconstructed scans can be reliably retrieved from

Algorithm 2 Metric-Guided Image Bound Retrieval and Outlier Detection.

Inputs: Test subject sample images $\hat{I}_{n+1}$ and metrics $\hat{Y}_{n+1}$, Prediction interval $C(\hat{Y}_{n+1}) = [L(\hat{Y}_{n+1}), U(\hat{Y}_{n+1})]$ from Alg. 1.
Outputs: Lower and upper bound reconstructions $L(\hat{I}_{n+1})$ and $U(\hat{I}_{n+1})$, inlier and outlier reconstructions $In(\hat{I}_{n+1})$ and $Out(\hat{I}_{n+1})$.

▷ Retrieve upper and lower bound reconstructions.
$L(\hat{I}_{n+1}) \leftarrow \arg\min_{\hat{I}_{n+1,j}} |\hat{Y}_{n+1,j} - L(\hat{Y}_{n+1})|$
$U(\hat{I}_{n+1}) \leftarrow \arg\min_{\hat{I}_{n+1,j}} |\hat{Y}_{n+1,j} - U(\hat{Y}_{n+1})|$

▷ Retrieve inliers and outlier reconstructions.
$In(\hat{I}_{n+1}) \leftarrow \{\hat{I}_{n+1,j} \in \hat{I}_{n+1} | \hat{Y}_{n+1,j} \in [L(\hat{Y}_{n+1}), U(\hat{Y}_{n+1})]\}$
$Out(\hat{I}_{n+1}) \leftarrow \{\hat{I}_{n+1,j} \in \hat{I}_{n+1} | \hat{Y}_{n+1,j} \notin [L(\hat{Y}_{n+1}), U(\hat{Y}_{n+1})]\}$

the metric bounds. Finally, we show that our framework flags dangerous outlier reconstructions that look plausible but have statistically unlikely metric values. Our work lays the foundation for more interpretable and trustworthy test-time assessments of medical image reconstruction algorithms.

2 Methods

We assume a fixed stochastic image acquisition system and reconstruction algorithm that outputs scans in $\mathbb{R}^{\Omega}$, and a given function $f(\cdot) : \mathbb{R}^{\Omega} \to \mathbb{R}^1$ that outputs a scalar metric of interest from a scan. We also assume an input calibration dataset $\mathcal{D} = \{\hat{Y}_i\}_{i=1}^{n}$ of n random subjects, where $\hat{Y}_i = \{\hat{Y}_{i,j}\}_{j=1}^{n_s}$ is a set of n_s predicted metrics (derived by running $f(\cdot)$ on n_s randomly sampled reconstructed scans). Let α_{lo} and α_{hi} denote desired lower and upper mis-coverage rates, and $Q_{\alpha}(\cdot)$ a function that returns the α-th quantile of a set.

For test subject $n+1$, we derive a prediction set $C(\hat{Y}_{n+1}) = [L(\hat{Y}_{n+1}), U(\hat{Y}_{n+1})]$ that yields marginal coverage: $P(Y_{n+1} \in C(\hat{Y}_{n+1})) \geq 1 - \alpha_{lo} - \alpha_{hi}$ [11]. If the reconstruction pipeline were perfectly calibrated, we could simply set $L(\hat{Y}_{n+1}) = Q_{\alpha_{lo}}(\hat{Y}_{n+1})$ and $U(\hat{Y}_{n+1}) = Q_{\alpha_{hi}}(\hat{Y}_{n+1})$. However,to handle mis-calibrations, we use $\mathcal{D}$ to find appropriate adjustments to $Q_{\alpha_{lo}}(\hat{Y}_{n+1})$ and $Q_{\alpha_{hi}}(\hat{Y}_{n+1})$. We find these adjustments in Sec. 2.1, and the use of the metric bounds to retrieve inlier, outlier, and nearest neighbor reconstructed scans in Sec. 2.2.

2.1 Quantile Adjustments Using Calibration Data

Using $\mathcal{D}$, we derive adjustments q_{lo} and q_{hi} to the $Q_{\alpha_{lo}}(\hat{Y}_{n+1})$ and $Q_{1-\alpha_{hi}}(\hat{Y}_{n+1})$ quantiles of the ground truth metric of test subject $n+1$ (Alg. 1). To do so, we use a probabilistic "split" CP setup proposed in previous work [32]. We first calculate $s_{lo,i}$ and $s_{hi,i}$, the lower and upper *non-conformity scores* for each patient, capturing the differences between the lower and upper sample quantiles and ground truth metrics: $s_{i,lo} = Q_{\alpha_{lo}}(\hat{Y}_i) - Y_i$ and $s_{i,hi} = Y_i - Q_{1-\alpha_{hi}}(\hat{Y}_i)$. We then

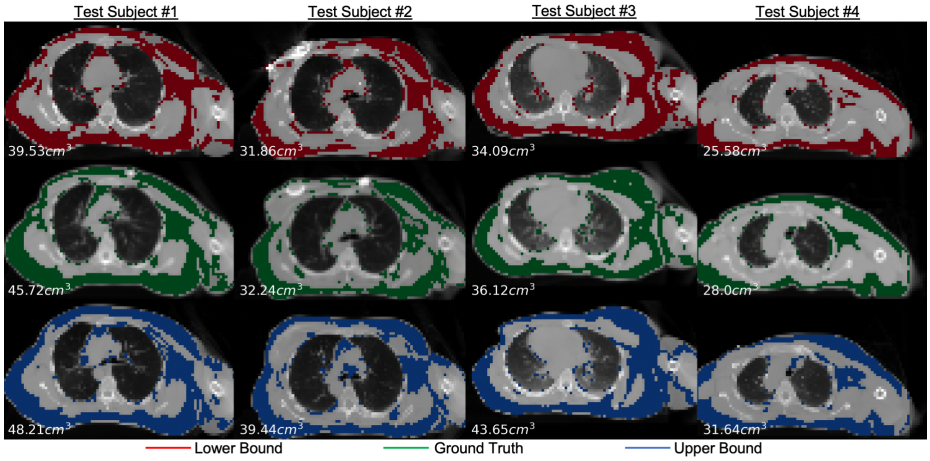

Fig. 2. Sample retrieved visual bounds for fat mass quantification. For four subjects, we show retrieved reconstructions closest to the metric lower bound (top row) and upper bound (bottom), along with the ground truth scan (middle). Per scan, we overlay predicted fat pixels and print total fat volume on the bottom left. The visual bounds provide context into which spatial regions significantly vary in the prediction interval, which may not be obvious from the metric bounds alone.

compute the $(1-\alpha_{lo})$-th and $(1-\alpha_{hi})$-th empirical quantiles of these scores to arrive at the adjustments: $q_{lo} = Q_{1-\hat{\alpha}_{lo}}(\{s_{i,lo}\}_{i=1}^{n})$ and $q_{hi} = Q_{1-\hat{\alpha}_{hi}}(\{s_{i,hi}\}_{i=1}^{n})$, where $\hat{\alpha}_{lo} = \lfloor \alpha_{lo}(n+1) \rfloor / n$ and $\hat{\alpha}_{hi} = \lfloor \alpha_{hi}(n+1) \rfloor / n$ are finite-sample adjusted mis-coverage rates.

At test time, we first reconstruct images $\hat{I}_{n+1} = \{\hat{I}_{n+1,j}\}_{j=1}^{n_s}$ for the test subject and compute the derived metrics $\hat{Y}_{n+1} = \{\hat{Y}_{n+1,j}\}_{j=1}^{n_s}$. We then simply compute the prediction interval $C(\hat{Y}_{n+1}) = [Q_{\alpha_{lo}}(\hat{Y}_{n+1}) - q_{lo}, Q_{1-\alpha_{hi}}(\hat{Y}_{n+1}) + q_{hi}]$ using the learned adjustments from $\mathcal{D}$. Under the assumption of exchangeability, this prediction interval is guaranteed to have marginal coverage [32].

2.2 Retrieving Scans Using Metric Bounds

While the metric bounds are useful in their own right to offer interpretable feedback on patient state, they can also be mapped back to the reconstructed scans $\{\hat{Y}_{n+1,j}\}_{j=1}^{n_s}$ to provide visual feedback to the user or clinician (Alg. 2). First, we can flag "outlier" scans, i.e., those scans with metrics outside $[L(\hat{I}_{n+1}), U(\hat{I}_{n+1})]$. Second, we can retrieve the scans with metrics nearest to $L(\hat{I}_{n+1})$ and $U(\hat{I}_{n+1})$, thereby providing visual interpretability to the metric bounds.

3 Experiments

We empirically evaluated our framework on sparse-view CT reconstruction applications: thoracic fat mass quantification (fat) and post-mastectomy radiother-

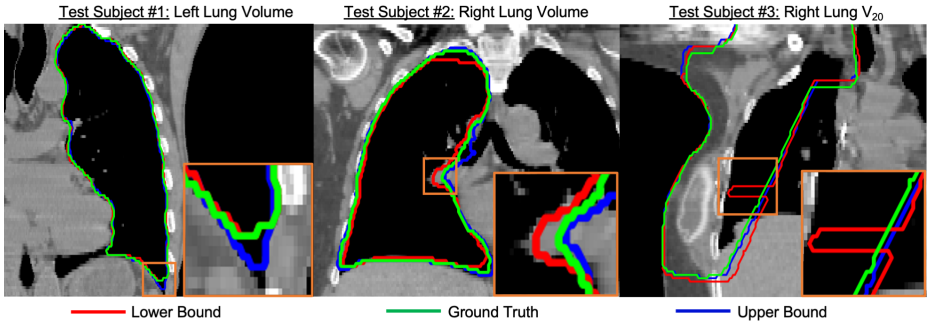

Fig. 3. Sample retrieved visual bounds for RT planning. We overlay contours on ground truth CT scans corresponding to RT metrics derived from: the ground truth scan (green), and reconstructed scans closest to the metric lower bound (red), and upper bound (blue). We show a different RT planning metric per subject. The bottom right inset of each scan zooms in on one region with significant contour differences. Left/right lung volume metrics are sensitive to reconstruction quality at small "corner"-like regions. The lower bound for V_{20} (subject 3) contains a significant spike, indicating sensitivity to some reconstructed artifact in that spatial region.

apy planning (RT), described below. We compared our approach, named *Metric CP*, to three baselines: *Metric*, *Pixel*, and *Pixel CP*. *Metric* simply computes the α_{lo}-th and $(1-\alpha_{hi})$-th quantiles of metrics across sampled images. *Pixel* computes pixel-wise α_{lo}-th and $(1-\alpha_{hi})$-th quantiles of intensity values across images [10, 13, 17]. *Pixel CP* applies CP to adjust the pixel-wise quantiles to achieve marginal coverage of ground-truth pixel intensities [26, 32]. For all experiments, we set $\alpha_{lo} = \alpha_{hi} = 0.05$ (target coverage of 0.9).

Fat Mass Estimation. Quantifying thoracic fat mass distribution is crucial to understanding its role in cardiometabolic risk and other health outcomes [7, 31]. We computed fat volume for a 2D CT slice by multiplying the number of fat pixels with slice thickness. We identified fat pixels by thresholding values in [-150,-50] Hu. We used a de-identified CT dataset of 935 subjects from The University of Texas M.D. Anderson Cancer Center. [3]. We split this dataset into 729 subjects (∼300k axial slices) for training and 265 subjects (795 axial slices) for testing. We trained an unconditional 2D U-Net diffusion model on training data and used guided posterior sampling to reconstruct slices from 15 1-D observed projections [8, 14]. We used $n = 596$ slices for calibration and 199 slices for testing, and generated $n_s = 100$ reconstructions per slice using random input noise initializations. We repeated the experiment 1000 times with random calibration-testing splits.

RT Planning. RT is a cornerstone of modern cancer treatment, and sparse-view CT can potentially lower equipment costs and improve access to CT imaging for RT planning [6]. We generated RT dose plans for CTs using the Radiation Planning Assistant (RPA, FDA 510(k) cleared) [1, 6, 19] and investigated sev-

[3] This research was conducted using an approved institutional review board protocol.

Table 1. Quantitative comparison of *Metric CP* to baselines. We used 1000 random 75%-25% calibration-testing splits, and report mean coverage and normalized interval lengths. The normalized interval length for a baseline method is its mean interval length divided by *Metric CP*'s mean interval length. While *Metric*, *Pixel*, and *Pixel CP* bounds often have smaller ("tighter") interval lengths than *Metric CP* (a normalized length < 1), they suffer from significant under-coverage, making them unusable for interpretable confidence assessments. By construction, *Metric CP* provides a target coverage with reasonable interval lengths.

		Test Coverage (↑)				Normalized Interval Length (↓)		
	Metric	Metric CP	Metric	Pixel CP	Pixel	Metric	Pixel CP	Pixel
Fat	Fat Mass	**0.902**	0.804	0.316	0.316	0.92	3.12	3.12
RT	Heart D_0	**0.932**	0.576	0.650	0.725	0.64	0.28	0.24
	Heart Vol	**0.935**	0.441	0.225	0.475	0.23	0.19	0.19
	Right Lung D_0	**0.941**	0.695	0.510	0.418	0.66	0.16	0.13
	Right Lung Vol	**0.930**	0.176	0.000	0.000	0.72	0.00	0.00
	Right Lung V_{20}	**0.934**	0.649	0.018	0.018	0.82	0.31	0.31
	Left Lung D_0	**0.958**	0.721	0.225	0.275	0.41	0.01	0.01
	Left Lung Vol	**0.935**	0.247	0.025	0.025	0.61	0.02	0.02

eral key dose and structural metrics: maximum dose to the heart/left lung/right lung (D_0), volumes of heart/left lung/right lung (Vol), and volume of right lung that received 20Gy (Right Lung V_{20}). We used a de-identified CT dataset of 40 subjects retrospectively treated with radiotherapy at The University of Texas M.D. Anderson Cancer Center. For each subject, we generated digitally reconstructed radiographs (DRRs) from the ground truth cone-beam CT at 50 uniformly spaced angles between 0° and 360°. We used a self-supervised reconstruction model, Neural Attenuation Fields (NAF) [34]. We used $n = 30$ scans for calibration and 10 scans for testing, and generated $n_s = 10$ reconstructions per scan using random initializations of network parameters. We repeated the experiment 1000 times with random calibration-testing splits.

Evaluation Metrics. Statistical prediction methods are faced with a trade-off between *test coverage* and *interval length* metrics. Test coverage is the probability that the ground truth metric for a test subject falls within the calibrated prediction interval. Interval length is the difference between upper and lower bounds (and 0 if the difference is negative). For baseline methods, we specifically report normalized interval length: interval lengths divided by *Metric CP*'s interval length, thereby removing arbitrary scaling differences across metrics.

3.1 Results

Table 1 presents quantitative results. *Metric CP* achieves valid coverage for each metric of choice, while other common baseline strategies suffer from significant under-coverage. Bounds derived from simply taking 0.05-th and 0.95-th quantiles of sampled metrics (without CP) or pixels therefore do not align well with calibrated metric prediction interval bounds. *Pixel* and *Pixel CP* yield small interval lengths, which may seem beneficial. However, they lead to poor coverage

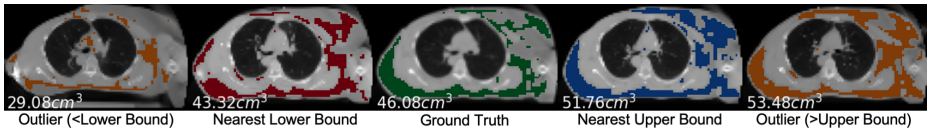

Fig. 4. Sample retrieved fat mass outliers for one subject. We show example outlier scans that have metrics outside the calibrated prediction intervals (columns 1 and 5), nearest neighbor lower and upper bound scans (columns 2 and 4), and the ground truth scan (column 3). Although outlier scans may look plausible, our framework flags these scans as not being statistically probable due to their metric values.

because they focus on pixel intensity, which is often not a key factor relevant to downstream metrics. *Metric* (without CP) still significantly outperforms both pixel approaches, showing that if a user wants a prediction interval for a given metric, it is better to simply estimate this interval using the metric rather than use any pixel-wise approach even with CP.

Fig. 2 and Fig. 3 present sample visual bound retrievals for fat estimation and RT planning metrics. We overlay per-pixel fat classification predictions on the scans in Fig. 2, and overlay RT planning contours in Fig. 3. The visual bounds for fat estimation in Fig. 2 provide context into which spatial regions vary significantly in fat content across the prediction interval. The visual bounds for RT planning in Fig. 3 demonstrate that often there are small regions of an organ that are responsible for the variance to RT planning values. Finally, Fig. 4 shows an example of detecting outliers from reconstructions for one subject for the fat estimation task. Although the outliers in this figure may look plausible, our framework flags these scans as not being statistically probable due to their extreme metric values. It is likely that the diffusion reconstruction algorithm may have some biases contributing to those cases.

4 Conclusion

We proposed a theoretically grounded, application-tuned framework to identify the quality of algorithmically reconstructed scans, and derive valid prediction intervals on derived metrics. Results show that representing scans with metrics, as opposed to operating in the raw pixel space, helps produce interpretable confidence intervals that capture meaningful spatial correlations between pixels [19, 1, 6]. However, naively computing confidence intervals directly from these metrics without *calibration* is insufficient, as they do not accurately reflect the true distribution of ground truth values. The proposed framework could prove useful in flagging situations where a scan-derived metric is statistically likely to be outside of acceptable safety bounds. For example, maximum dose to the heart in RT planning should be <45Gy. If bounds for a subject indicate a good chance of the metric going beyond this limit, the case can be flagged for further review and intervention [21, 25, 33].

Our framework has a few assumptions and limitations to consider in practice. First, CP assumes exchangeability of data samples. Practitioners should ensure that calibration data is representative of future test data. Furthermore, the current method focuses on computing bounds on one metric, but there are often several metrics of interest in any application. A natural extension of this work would be to derive multi-metric bounds for image reconstruction. Moreover, image bound retrieval is only useful if there are reconstructions in the scan set with metric values sufficiently close to the metric bounds. If, for example, the reconstruction algorithm has significant bias or the number of samples n_s is small, this may not hold and lead to misleading visual interpretations. Finally, more case studies are warranted to demonstrate the utility of these visual bounds.

Acknowledgments. The authors would also like to acknowledge support from a fellowship from the Gulf Coast Consortia on the NLM Training Program in Biomedical Informatics and Data Science T15LM007093. The authors would also like to thank the RPA team (Joy Zhang, Raphael Douglas) for their support. Tucker Netherton would like to acknowledge the support of the NIH LRP award.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aggarwal, A., Burger, H., Cardenas, C., Chung, C., Douglas, R., du Toit, M., Jhingran, A., Mumme, R., Muya, S., Naidoo, K., et al.: Radiation planning assistant-a web-based tool to support high-quality radiotherapy in clinics with limited resources. Journal of Visualized Experiments: Jove (200) (2023)
2. Angelopoulos, A.N., Kohli, A.P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., Romano, Y.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: International Conference on Machine Learning. pp. 717–730. PMLR (2022)
3. Belhasin, O., Romano, Y., Freedman, D., Rivlin, E., Elad, M.: Principal uncertainty quantification with spatial correlation for image restoration problems. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
4. Bhadra, S., Kelkar, V.A., Brooks, F.J., Anastasio, M.A.: On hallucinations in tomographic image reconstruction. IEEE transactions on medical imaging **40**(11), 3249–3260 (2021)
5. Chan, M.A., Molina, M.J., Metzler, C.: Estimating epistemic and aleatoric uncertainty with a single model. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
6. Court, L., Aggarwal, A., Burger, H., Cardenas, C., Chung, C., Douglas, R., du Toit, M., Jaffray, D., Jhingran, A., Mejia, M., et al.: Addressing the global expertise gap in radiation oncology: the radiation planning assistant. JCO Global Oncology **9**, e2200431 (2023)
7. Dey, D., Nakazato, R., Li, D., Berman, D.S.: Epicardial and thoracic fatnoninvasive measurement and clinical implications. Cardiovascular diagnosis and therapy **2**(2), 85 (2012)

8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Diwakar, M., Kumar, M.: A review on ct image noise and its denoising. Biomedical Signal Processing and Control **42**, 73–88 (2018)
10. Edupuganti, V., Mardani, M., Vasanawala, S., Pauly, J.: Uncertainty quantification in deep mri reconstruction. IEEE Transactions on Medical Imaging **40**(1), 239–250 (2020)
11. Fontana, M., Zeni, G., Vantini, S.: Conformal prediction: a unified review of theory and new challenges. Bernoulli **29**(1), 1–23 (2023)
12. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(Suppl 1), 1513–1589 (2023)
13. Gillmann, C., Saur, D., Wischgoll, T., Scheuermann, G.: Uncertainty-aware visualization in medical imaging-a survey. In: Computer Graphics Forum. vol. 40, pp. 665–689. Wiley Online Library (2021)
14. Graikos, A., Malkin, N., Jojic, N., Samaras, D.: Diffusion models as plug-and-play priors. Advances in Neural Information Processing Systems **35**, 14715–14728 (2022)
15. Herman, G.T.: Fundamentals of computerized tomography: image reconstruction from projections. Springer Science & Business Media (2009)
16. Horwitz, E., Hoshen, Y.: Conffusion: Confidence intervals for diffusion models. arXiv preprint arXiv:2211.09795 (2022)
17. Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A.G., Tamir, J.: Robust compressed sensing mri with deep generative priors. Advances in Neural Information Processing Systems **34**, 14938–14954 (2021)
18. Kalisz, K., Buethe, J., Saboo, S.S., Abbara, S., Halliburton, S., Rajiah, P.: Artifacts at cardiac ct: physics and solutions. Radiographics **36**(7), 2064–2083 (2016)
19. Kisling, K., McCarroll, R., Zhang, L., Yang, J., Simonds, H., Du Toit, M., Trauernicht, C., Burger, H., Parkes, J., Mejia, M., et al.: Radiation planning assistant-a streamlined, fully automated radiotherapy treatment planning system. JoVE (Journal of Visualized Experiments) (134), e57411 (2018)
20. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
21. Lekeufack, J., Angelopoulos, A.A., Bajcsy, A., Jordan, M.I., Malik, J.: Conformal decision theory: Safe autonomous decisions from imperfect predictions. arXiv preprint arXiv:2310.05921 (2023)
22. Lustig, M., Donoho, D.L., Santos, J.M., Pauly, J.M.: Compressed sensing mri. IEEE signal processing magazine **25**(2), 72–82 (2008)
23. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
24. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: 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)
25. Romano, Y., Barber, R.F., Sabatti, C., Candès, E.: With malice toward none: Assessing uncertainty via equalized coverage. Harvard Data Science Review **2**(2), 4 (2020)
26. Romano, Y., Patterson, E., Candes, E.: Conformalized quantile regression. Advances in neural information processing systems **32** (2019)

27. Sankaranarayanan, S., Angelopoulos, A., Bates, S., Romano, Y., Isola, P.: Semantic uncertainty intervals for disentangled latent spaces. Advances in Neural Information Processing Systems **35**, 6250–6263 (2022)
28. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)
29. Teneggi, J., Tivnan, M., Stayman, W., Sulam, J.: How to trust your diffusion model: A convex optimization approach to conformal risk control. In: International Conference on Machine Learning. pp. 33940–33960. PMLR (2023)
30. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world, vol. 29. Springer (2005)
31. Wang, H., Chen, Y., Eitzman, D.T.: Imaging body fat: techniques and cardiometabolic implications. Arteriosclerosis, thrombosis, and vascular biology **34**(10), 2217–2223 (2014)
32. Wang, Z., Gao, R., Yin, M., Zhou, M., Blei, D.M.: Probabilistic conformal prediction using conditional random samples. arXiv preprint arXiv:2206.06584 (2022)
33. Ye, C.T., Han, J., Liu, K., Angelopoulos, A., Griffith, L., Monakhova, K., You, S.: Learned, uncertainty-driven adaptive acquisition for photon-efficient multiphoton microscopy. arXiv preprint arXiv:2310.16102 (2023)
34. Zha, R., Zhang, Y., Li, H.: Naf: neural attenuation fields for sparse-view cbct reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 442–452. Springer (2022)