Improving Trustworthiness of AI Disease Severity Rating in Medical Imaging with Ordinal Conformal Prediction Sets

Charles Lu*, Anastasios N. Angelopoulos*, Stuart Pomerantz

Abstract

The regulatory approval and broad clinical deployment of medical AI have been hampered by the perception that deep learning models fail in unpredictable and possibly catastrophic ways. A lack of statistically rigorous uncertainty quantification is a significant factor undermining trust in AI results. Recent developments in distribution-free uncertainty quantification present practical solutions for these issues by providing reliability guarantees for black-box models on arbitrary data distributions as formally valid finite-sample prediction intervals. Our work applies these new uncertainty quantification methods — specifically conformal prediction — to a deep-learning model for grading the severity of spinal stenosis in lumbar spine MRI. We demonstrate a technique for forming ordinal prediction sets that are guaranteed to contain the correct stenosis severity within a user-defined probability (confidence interval). On a dataset of 409 MRI exams processed by the deep-learning model, the conformal method provides tight coverage with small prediction set sizes. Furthermore, we explore the potential clinical applicability of flagging cases with high uncertainty predictions (large prediction sets) by quantifying an increase in the prevalence of significant imaging abnormalities (e.g. motion artifacts, metallic artifacts, and tumors) that could degrade confidence in predictive performance when compared to a random sample of cases.

1 Introduction

Although many studies have demonstrated high overall accuracy in automating medical imaging diagnosis with deep-learning AI models, translation to actual clinical deployment has proved difficult. It is widely observed that deep learning algorithms can fail in bizarre ways and with misplaced confidence [23, 14]. A core problem is a lack of *trust* — a survey of radiologists found they that although they thought AI tools add value to their clinical practice, they would not trust AI for autonomous clinical use due to perceived and experienced unreliability [1].

*Equal contribution

Herein, we present methods for endowing arbitrary AI systems with formal mathematical guarantees that give clinicians explicit assurances about an algorithm's overall performance and most importantly for a given study. These guarantees are *distribution-free*—they work for any (pre-trained) model, any (possibly unknown) data distribution, and in finite samples. Although such guarantees do not solve the issue of trust entirely, the precise and formal understanding of their model's predictive uncertainty enables a clinician to potentially work more assuredly in concert with AI assistance.

We demonstrate the utility of our methods using an AI system developed to assist radiologists in the grading of spinal stenosis in lumbar MRI. Degenerative spinal stenosis is the abnormal narrowing of the spinal canal that compresses the spinal cord or nerve roots, often resulting in debility from pain, limb weakness, and other neurological disorders. It is a highly prevalent condition that affects working-age and elderly populations and constitutes a heavy societal burden not only in the form of costly medical care but from decreased workplace productivity, disability, and lowered quality of life. The formal interpretation of spinal stenosis imaging remains a challenging and time-consuming task even for experienced subspecialty radiologists due to the complexity of spinal anatomy, pathology, and the MR imaging modality. Using a highly accurate AI model to help assess the severity of spinal stenosis on MRI could lower interpretation time and improve the consistency of grading. [12] Yet, given the practical challenges of medical imaging that can degrade model performance in any given exam, the adoption of such tools will be low if clinicians encounter poor quality predictions without a sense of when the model is more or less reliable. To bolster such trust, we apply conformal prediction to algorithmic disease severity classification sets in order to identify higher uncertainty predictions that might merit special attention by the radiologist. Our main contributions are the following:

1. We develop new distribution-free uncertainty quantification methods for ordinal labels.

2. To our knowledge, we are the first to apply distribution-free uncertainty quantification to the results of an AI model for automated stenosis grading of lumbar spinal MRI.

3. We identify a correlation between high prediction uncertainty in individual cases and the presence of potentially contributory imaging features as evaluated by a neuroradiologist such as tumors, orthopedic hardware artifacts, and motion artifacts.

2 Methods

To formally describe the problem, let the input $X_{\text{test}} \in \mathcal{X}$, $\mathcal{X} = \mathbb{R}^{H \times W \times D}$ be an MR image and the ground truth label $Y_{\text{test}} \in \mathcal{Y}$, $\mathcal{Y} = \{0, ..., K-1\}$ be an ordinal value representing the severity of the disease (higher values indicating greater severity). We are given a pre-trained model, $\hat{f}$, that takes in images and

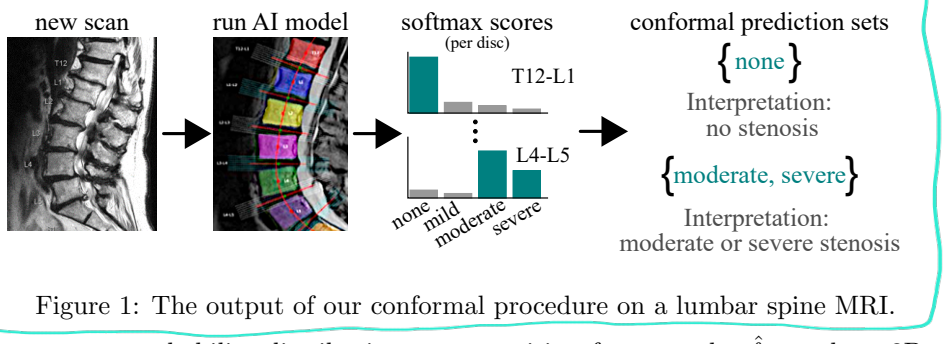

Figure 1: The output of our conformal procedure on a lumbar spine MRI.

outputs a probability distribution over severities; for example, $\hat{f}$ may be a 3D convolutional neural network with a softmax function. Assume we also have a calibration dataset, $\{(X_i, Y_i)\}_{i=1}^{n}$, of data points that the model has not seen before. This calibration data should be composed of pairs of MR images and their associated disease severity labels drawn i.i.d. Given a new MR image X_{test}, the task is to predict the (unobserved) severity, Y_{test}. In the usual multi-class classification setting, the output is the label with the highest estimated probability, $\hat{Y}(x) = \underset{y \in \{1,\ldots,K\}}{\arg\max}\, \hat{f}(x)_y$. However, $\hat{Y}(X_{\text{test}})$ may be wrong, either because the learned model $\hat{f}$ does not learn the relationship between MR images and severities properly or because there is intrinsic randomness in this relationship that cannot be accounted for by any algorithm (i.e. aleatoric uncertainty).

Our goal is to rigorously quantify this uncertainty by outputting a set of probable disease severities that is guaranteed to contain the ground truth severity on average. These *prediction sets* will provide *distribution-free* probabilistic guarantees, *i.e.* ones that do not depend on the model or distribution of the data.

Ordinal Adaptive Prediction Sets (Ordinal APS)

Our approach uses conformal prediction with a novel score function designed for ordinal labels. In particular, each prediction set will always be a contiguous set of severities, and for any user-specified error rate α, prediction sets will contain the true label with probability $1 - \alpha$. The reader can refer to [25] and [27] for similar algorithms and exposition.

An Oracle Method

Imagine we had oracle access to the true probability distribution over severities $P(Y_{\text{test}} \mid X_{\text{test}})$ with associated density function $f(x)_y$. A reasonable goal might then be to pick the set with the smallest size while still achieving *conditional coverage*.

Definition 1 (conditional coverage)**.** *A predicted set of severities $\mathcal{T}(X_{\text{test}})$ has conditional coverage if it contains the true severity with $1 - \alpha$ probability no*

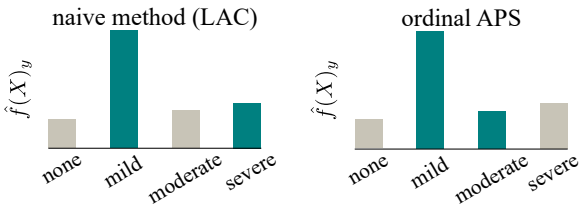

Figure 2: Comparison of an example prediction set chosen by naive least-ambiguous set-valued (LAC) classifiers [26] and Ordinal APS methods; notice that LAC does not respect ordinality, which may result in an incongruous prediction set.

matter what MR image is input, i.e.,

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{T}(x) \mid X_{\text{test}} = x\right) \geq 1 - \alpha, \text{ for all } x \in \mathcal{X}. \tag{1}$$

The clinical benefit of conditional coverage is that it essentially achieves a per-patient guarantee as opposed to one that is averaged across patients. Under conditional coverage, the uncertainty sets will work equally well for all possible subgroups such as subpopulations from different patient demographics.

Ignoring tie-breaking, we can succinctly describe the oracle prediction set as follows:

$$\mathcal{T}^{(\text{optimal})}(x) = [l^*(x),\ u^*(x)]\,, \text{ where}$$
$$\left(l^*(x),\ u^*(x)\right) = \underset{\substack{(l,u)\in\mathcal{Y}^2 \\ l\leq u}}{\arg\min}\left\{u - l : \sum_{j=l}^{u} f(x)_j \geq 1 - \alpha\right\}. \tag{2}$$

This set, $\mathcal{T}^{(\text{optimal})}$, is the smallest that satisfies (1). Ideally, we would compute $\mathcal{T}^{(\text{optimal})}$ exactly, but we do not have access to f, only its estimator $\hat{f}$, which may be arbitrarily bad.

Ordinal Adaptive Prediction Sets

Naturally, the next step is to plug in our estimate of the probabilities, $\hat{f}$, to (2). However, because $\hat{f}$ may be wrong, we must calibrate the resulting set with conformal prediction, yielding a *marginal coverage* guarantee. Our procedure is illustrated graphically in the right plot of Figure 2; it corresponds to greedily growing the set outwards from the maximally likely predicted severity (i.e. "mild" stenosis in this example).

Definition 2 (marginal coverage)**.** *A predicted set of severities $\mathcal{T}$ has marginal coverage if it contains the true severity on average over new MRIs, i.e.,*

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{T}(X_{\text{test}})\right) \geq 1 - \alpha. \tag{3}$$

Marginal coverage is weaker than conditional coverage since it holds only on average over the entire population, so coverage may be worse or better for some subgroups. While conditional coverage is, in general, impossible [7], we can

hope to approximate conditional coverage by defining a set similar to $\mathcal{T}^{(\text{optimal})}$ that uses $\hat{f}$. To that end, define a sequence of sets indexed by a threshold λ,

$$\mathcal{T}_\lambda(x) = [\hat{l}^*(x),\ \hat{u}^*(x)],\ \text{where}$$
$$\left(\hat{l}^*(x),\ \hat{u}^*(x)\right) = \underset{\substack{(l,\,u)\,\in\,\mathcal{Y}^2 \\ l\,\leq\,u}}{\arg\min} \left\{ u - l : \sum_{j=l}^{u} \hat{f}(j|x) \geq \lambda \right\}.$$

Notice that as λ grows, the sets grow, meaning they are *nested* in λ:

$$\lambda_1 \leq \lambda_2 \implies \forall x,\ \ \mathcal{T}_{\lambda_1}(x) \subseteq \mathcal{T}_{\lambda_2}(x).$$

The key is to pick a value of λ such that the resulting set satisfies (3). The following algorithm takes as input $\mathcal{T}_\lambda$ and outputs our choice of λ:

$$\mathcal{A}(\mathcal{T}_\lambda;\ \alpha) = \inf \left\{ \lambda : \sum_{i=1}^{n} \mathbb{1}\left\{Y_i \in \mathcal{T}_\lambda(X_i)\right\} \ \geq\ \lceil (n+1)(1-\alpha) \rceil \right\}.$$

The key to our guarantee is the quantity $\lceil (n+1)(1-\alpha) \rceil$, which is slightly larger than the naive choice $n(1-\alpha)$ and helps us correct for the model's deficiencies; see [2] for details on this statistical argument. Using this algorithm, approximated in Algorithm 1, results in a marginal coverage guarantee.

Theorem 1 (Conformal coverage guarantee)**.** *Let* (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) *and* $(X_{\text{test}}, Y_{\text{test}})$ *be an i.i.d. sequence of MRIs and paired severities and let* $\hat{\lambda} = \mathcal{A}(\mathcal{T}_\lambda,\ \alpha)$*. Then* $\mathcal{T}_{\hat{\lambda}}$ *satisfies* (3)*, i.e., it contains the true label with probability* $1 - \alpha$*.*

This theorem holds for any data distribution or machine learning model, any number of calibration data points n, and any possible sequence of nested sets that includes $\mathcal{Y}$ (see the formal version and proof in Appendix A).

Implementing Ordinal Adaptive Prediction Sets

In practice, Ordinal APS has two undesirable properties: computing $\mathcal{T}_\lambda(x)$ requires a combinatorial search of the set of possible severities, and $\mathcal{T}_\lambda(x)$ may not include the point prediction $\hat{Y}$. In practice, we therefore approximate Ordinal APS greedily as described below in Algorithm 1.

The algorithm always contains $\hat{Y}$ and requires only $\mathcal{O}(n)$ computations; furthermore, it usually results in exactly the same sets as the exact method in our experiments, which have a small value of K.

Note that the approximate choice of $\mathcal{T}_\lambda(x)$ described in Algorithm 1 is still nested, and thus we can still guarantee coverage (see Corollary A.1 for a formal statement and proof).

3 Experiments

We compare Ordinal Adaptive Prediction Sets to two other conformal methods: *Least Ambiguous set-valued Classifier* (LAC) [26] and *Ordinal Cumulative Distribution Function* (CDF). LAC uses the softmax score of the true class as the

Algorithm 1 Pseudocode for approximately computing $\mathcal{T}_\lambda(x)$

Input: Parameter λ; underlying predictor $\hat{f}$; input $x \in \mathcal{X}$.
Output: $\mathcal{T}_\lambda(x)$.

$\mathcal{T}_\lambda(x) \leftarrow \arg\max \hat{f}(x)$
$q \leftarrow 0$
while $q \leq \lambda$ **do**
  $S \leftarrow \{\min \mathcal{T}_\lambda(x) - 1,\ \max \mathcal{T}_\lambda(x) + 1\}$
  $y \leftarrow \underset{y' \in S}{\arg\max}\, \hat{f}(x)\mathbb{1}\{y' \in \{1, ..., K\}\}$
  $q \leftarrow q + \hat{f}(x)_y$
  $\mathcal{T}_\lambda(x) \leftarrow \mathcal{T}_\lambda(x) \cup \{y\}$

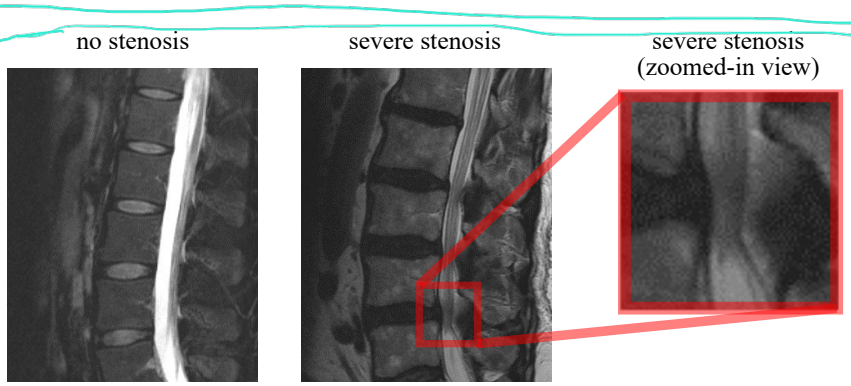

Figure 3: Example of a case without lumbar spinal stenosis and with severe lumbar spinal stenosis.

conformal score function. LAC theoretically gives the smallest average set size but sacrifices conditional coverage to achieve this. Additionally, LAC does not respect ordinality and thus may output non-contiguous prediction sets, which are inappropriate in an ordinal setting such as disease severity rating. The Ordinal CDF method starts at the highest prediction score and then inflates the intervals by λ in quantile-space; in that sense, it is similar to a discrete version of conformalized quantile regression [15, 24]. We only use the non-randomized versions of these methods.

We evaluate these three conformal methods on a deep learning system previously developed for automated lumbar spine stenosis grading in MRI, DeepSPINE [21]. The deep learning system consists of two convolutional neural networks – one to segment out and label each vertebral body and disc-interspace and the other to perform multi-class ordinal stenosis classification for three different anatomical sites (the central canal and right and left neuroforamina) at each intervertebral disc level for a total of up to 18 gradings per patient.

For each MRI exam, the associated radiology report was automatically parsed for keywords indicative of the presence and severity of stenosis for each of the 6 vertebral disc levels (T12-L1 through L5-S1) to extract ground truth labels for a total of 6,093 gradings. Each grading was assigned to a value on a

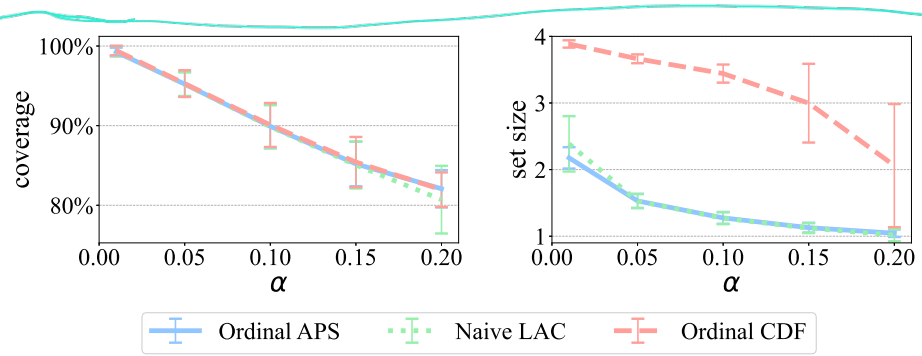

Figure 4: Empirical coverage and set size for three conformal prediction methods for $\alpha \in \{0.2, 0.15, 0.1, 0.05, 0.01\}$ (averaged over 100 trials and shown with $\pm$ 1 standard deviation).

four-point ordinal scale of stenosis severity: 0 (*no stenosis*), 1 (*mild stenosis*), 2 (*moderate stenosis*), and 3 (*severe stenosis*). Examples of patients with and without stenosis are shown in Figure 3.

For our experiments, we treat the stenosis grading model as a static model and only use it for inference to make predictions. We then process these predicted scores with the split-conformal techniques described in Section 2. This scenario would most closely reflect the clinical reality of incorporating regulated, third-party AI software medical devices, which would likely not permit users access to the AI model beyond the ability to make predictions.

Our code and analysis used in the quantitative experiments are made available here: `https://github.com/clu5/lumbar-conformal`.

3.1 Quantitative Experiments

We use the predicted softmax scores from a held-out set of MRI exams from 409 patients, comprising 6,093 disc level stenosis severity predictions to calibrate and evaluate each conformal methods. We randomly include 5% of patients in the calibration set and reserve the remainder for evaluating coverage and set size. We evaluate performance at several different α thresholds, $\alpha \in \{0.2, 0.15, 0.1, 0.05, 0.01\}$, and average results over 100 random trials.

As expected, all three conformal methods empirically achieve the desired marginal coverage as guaranteed by Theorem 1. However, Ordinal CDF requires a much larger set size to attain proper coverage than either Naive LAC or Ordinal APS for all values of α (shown in Figure 4).

In addition, while aggregate coverage is satisfied for each method, we find significant differences in class-conditional coverage (i.e. prediction sets stratified by the true stenosis severity label), which is shown in Figure 5. We see that prediction sets for "mild stenosis" and "moderate stenosis" grades have lower average coverage and larger set sizes than prediction sets for "no stenosis" and "severe stenosis" grades. These differences may be partly attributed to the fact that the "no stenosis" class constitutes the majority of the label

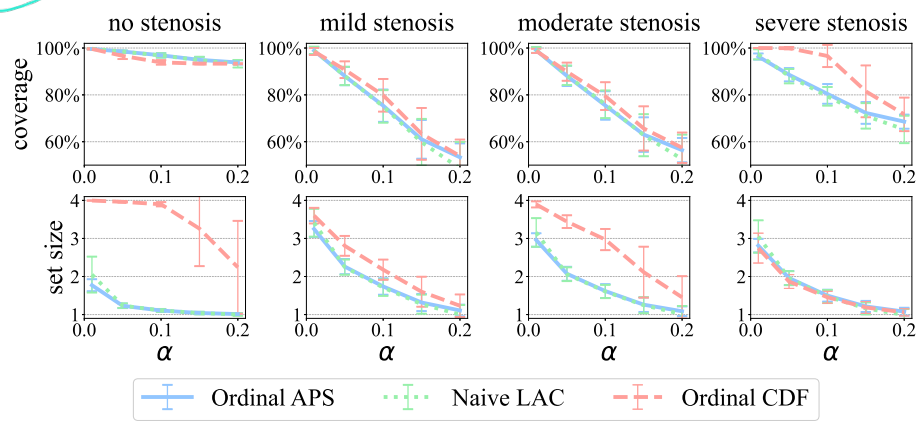

Figure 5: Comparing coverage and set size when stratified by ground-truth stenosis severity grading.

distribution (67%) and so may be easier for a model trained on this distribution to classify. Additionally, "mild stenosis" and "moderate stenosis" grades may be more challenging to differentiate than "no stenosis" and "severe stenosis" grades, reflecting the greater inherent variability and uncertainty in the ground truth ratings.

We also compare coverage and distribution stratified by set size in Figure 6. Stratifying by set size reveals that most predictions with Ordinal APS and Naive LAC contain only one or two grading predictions while Ordinal CDF mainly predicts a set of all four possible gradings (which always trivially satisfies coverage).

Lastly, we compare coverage and set size at the disc level in Table 1 at $\alpha = 0.1$. We find that coverage was inversely correlated to the prevalence of severe stenosis, which is most often found in the lower lumbar disc levels.

Overall, we conclude that Ordinal APS performs similarly to LAC in both coverage and set size, and both Ordinal APS and Naive LAC outperform Ordinal CDF. The similarities between LAC and Ordinal APS are notable — they almost always result in the same sets, although the algorithms are quite different. This is unsurprising given that in our setting $|\mathcal{Y}| = 4$ and the model's accuracy is high, so bimodal softmax scores almost never happen. Therefore LAC and Ordinal APS do the same thing. This observation does not generalize; other ordinal prediction problems with more categories will have bimodal distributions and thus LAC and Ordinal APS will differ.

3.2 Clinical Review of High Uncertainty Predictions

To investigate the clinical utility of Ordinal APS to enhance AI-augmented workflows, we evaluate one possible clinical integration use case: flagging low confidence predictions (i.e. ones with a large set size). The radiologist's performance and user experience of the model may be improved by raising their awareness

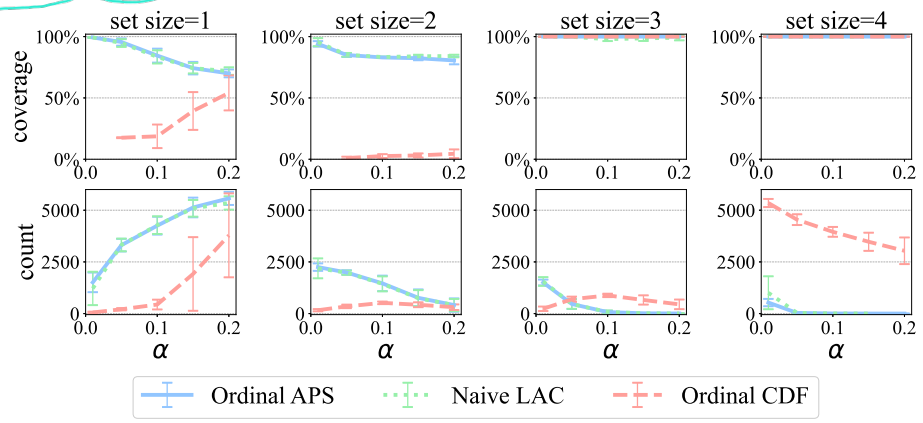

Figure 6: Comparing coverage and count of number of sets with a particular size when stratified by prediction set size (coverage is only shown if there was at least one prediction set at the desired α for a particular method).

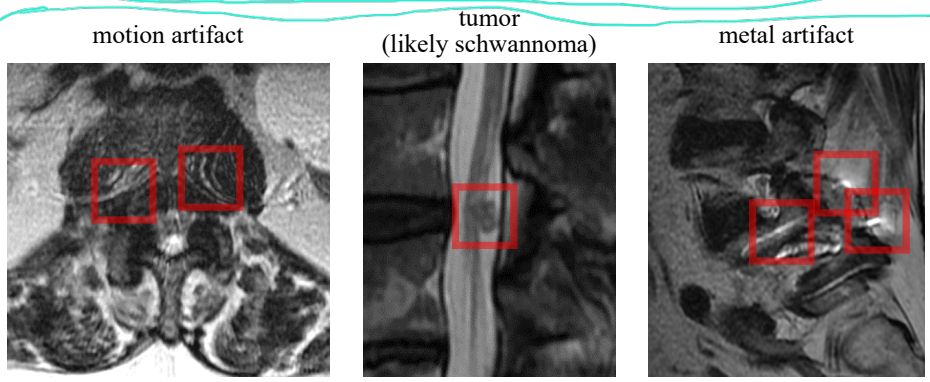

Figure 7: Three anomalies found in high uncertainty predictions; the anomalous areas are boxed in red.

of those instances in which the model performance may be degraded by scan anomalies or when uncertainty quantification is very high, excluding such potentially poor quality results from their review responsibilities altogether. We define an uncertainty score for each patient by taking the average set size for all disc levels and grading tasks. An neuroradiologist with > 20 years of experience determined what constituted a "significant imaging anomaly" within the context of spinal stenosis interpretation.

As a statistical validation of these results, we examined the report of 70 cases with the highest uncertainty and found 17 such anomalies: 11 cases with artifacts from metallic orthopedic hardware, four cases with motion artifacts, one case with a large tumor occupying the spinal canal, and one case with a severe congenital abnormality (achondroplastic dwarfism). In contrast, a random sample of 70 cases from the dataset only demonstrated five cases with significant

Table 1: Coverage and set size stratified by intervertebral disc level at $\alpha = 0.1$.

disc level	average severity	method	coverage	set size
T12-L1	0.04	Ordinal CDF	$99.7\% \pm 0.3\%$	3.97 ± 0.02
		Naive LAC	$98.6\% \pm 0.4\%$	1.04 ± 0.02
		Ordinal APS	$98.6\% \pm 0.3\%$	1.04 ± 0.01
L1-L2	0.18	Ordinal CDF	$97.5\% \pm 1.1\%$	3.84 ± 0.06
		Naive LAC	$95.4\% \pm 1.0\%$	1.12 ± 0.04
		Ordinal APS	$95.4\% \pm 1.0\%$	1.12 ± 0.04
L2-L3	0.48	Ordinal CDF	$90.4\% \pm 2.1\%$	3.50 ± 0.11
		Naive LAC	$90.0\% \pm 3.0\%$	1.25 ± 0.09
		Ordinal APS	$90.1\% \pm 3.0\%$	1.26 ± 0.09
L3-L4	0.75	Ordinal CDF	$84.3\% \pm 4.0\%$	3.17 ± 0.19
		Naive LAC	$86.0\% \pm 4.0\%$	1.42 ± 0.12
		Ordinal APS	$85.7\% \pm 4.2\%$	1.42 ± 0.13
L4-L5	1.06	Ordinal CDF	$81.7\% \pm 4.2\%$	2.86 ± 0.21
		Naive LAC	$83.2\% \pm 4.6\%$	1.48 ± 0.15
		Ordinal APS	$83.2\% \pm 4.8\%$	1.48 ± 0.15
L5-S1	0.71	Ordinal CDF	$84.4\% \pm 4.1$	3.20 ± 0.17
		Naive LAC	$85.3\% \pm 3.2$	1.39 ± 0.12
		Ordinal APS	$85.7\% \pm 3.0$	1.41 ± 0.11
Total	0.54	Ordinal CDF	$90.0\% \pm 2.5\%$	3.44 ± 0.12
		Naive LAC	$90.0\% \pm 2.6\%$	1.28 ± 0.09
		Ordinal APS	$90.1\% \pm 2.7\%$	1.28 ± 0.09

anomalies which were all orthopedic hardware artifacts. This difference is significant with $p < 0.05$ by Fisher's exact test, and qualitatively the abnormalities found in the filtered samples were more extreme.

Our manual review of high uncertainty cases shows promise for improving the clinician experience with AI-assisted medical software tools using distribution-free uncertainty quantification. Rather than presenting all AI predictions as equally trustworthy, cases with higher uncertainty can be flagged to prompt additional scrutiny or hidden from the user altogether to maximize efficiency. While prospective evaluation of this use of conformal prediction in more clinically realistic settings will be needed to validate general feasibility and utility, our preliminary experiments are a step towards demonstrating the clinical applicability of conformal prediction.

4 Related Work

Conformal Prediction and Distribution-Free Uncertainty Quantification

Conformal prediction is a flexible technique for generating prediction intervals from arbitrary models. It was first developed by Vladimir Vovk and collabo-

rators in the late 1990s [32, 31, 17, 18, 16, 26]. We build most directly on the work of Yaniv Romano and collaborators, who developed the Adaptive Prediction Sets method studied in [25, 6] and the histogram binning method in [27]. The latter work is the most relevant, and it proposes an algorithm very similar to Algorithm 1 in a continuous setting with histogram regression. Our work also relies on the nested set outlook on conformal prediction [13]. We also build directly on existing work involving distribution-free risk-controlling prediction sets and Learn then Test [8, 3]. The LAC baseline is taken from [26], and the ordinal CDF baseline is similar to the softmax method in [5], which is in turn motivated by [15, 24]. A gentle introduction to these topics and their history is available in [2], or alternatively, in [28].

Uncertainty Quantification for Critical Applications

Recently, uncertainty quantification has been promoted to facilitate trustworthiness and transparency in black-box algorithms, such as deep learning, for critical decision-making [10]. In particular, conformal prediction methods have been applied to a wide range of safety-critical applications – from reducing false alarms in the detection of sepsis risk [29] to end-to-end autonomous driving control [22]. Distribution-free uncertainty quantification techniques such as conformal prediction sets have emerged as an essential tool for rigorous statistical guarantees in medical decision-making [9, 20, 4, 30, 11, 19].

5 Conclusion

We show how conformal prediction sets can complement existing AI systems without further modification to the model to provide distribution-free reliability guarantees. We demonstrate its clinically utility in the application of flagging high uncertainty cases in automated stenosis severity grading for followup review. We hope this work promotes further studies on the trustworthiness and usability of uncertainty-aware machine learning systems for clinical applications.

Acknowledgements

SP receives research support from GE Healthcare. AA is funded by the NSF-GRFP and a Berkeley Fellowship. We thank Stephen Bates for many helpful conversations.

References

[1] Allen, B., Agarwal, S., Coombs, L., Wald, C., Dreyer, K.: 2020 acr data science institute artificial intelligence survey. Journal of the American College of Radiology **18**(8), 1153–1159 (2021)

[2] Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

[3] Angelopoulos, A.N., Bates, S., Candès, E.J., Jordan, M.I., Lei, L.: Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052 (2021)

[4] Angelopoulos, A.N., Bates, S., Zrnic, T., Jordan, M.I.: Private prediction sets. arXiv preprint arXiv:2102.06202 (2021)

[5] Angelopoulos, A.N., Kohli, A.P., Bates, S., Jordan, M.I., Malik, J., Alshaabi, T., Upadhyayula, S., Romano, Y.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. arXiv preprint arXiv:2202.05265 (2022)

[6] Angelopoulos, A.N., Bates, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction. In: International Conference on Learning Representations (ICLR) (2021)

[7] Barber, R., Candès, E., Ramdas, A., Tibshirani, R.: The limits of distribution-free conditional predictive inference. Information and Inference **10**(2), 455–482 (08 2021)

[8] Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. Journal of the Association for Computing Machinery **68**(6) (9 2021)

[9] Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature (01 2019)

[10] Bhatt, U., Antorán, J., Zhang, Y., Liao, Q.V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., Xiang, A.: Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty, p. 401–413. Association for Computing Machinery, New York, NY, USA (2021)

[11] Fannjiang, C., Bates, S., Angelopoulos, A., Listgarten, J., Jordan, M.I.: Conformal prediction for the design problem. arXiv preprint arXiv:2202.03613 (2022)

[12] Fu, M., Buerba, R., III, W., Blizzard, D., Lischuk, A., Haims, A., Grauer, J.: Inter-rater and intra-rater agreement of magnetic resonance imaging findings in the lumbar spine: Significant variability across degenerative conditions. The Spine Journal **14** (10 2014). https://doi.org/10.1016/j.spinee.2014.03.010

[13] Gupta, C., Kuchibhotla, A.K., Ramdas, A.: Nested conformal prediction and quantile out-of-bag ensemble methods. Pattern Recognition p. 108496 (2021)

[14] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. CVPR (2021)

[15] Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica: Journal of the Econometric Society **46**(1), 33–50 (1978)

[16] Lei, J.: Classification with confidence. Biometrika **101**(4), 755–769 (10 2014)

[17] Lei, J., Rinaldo, A., Wasserman, L.: A conformal prediction approach to explore functional data. Annals of Mathematics and Artificial Intelligence **74**, 29–43 (2015)

[18] Lei, J., Robins, J., Wasserman, L.: Distribution-free prediction sets. Journal of the American Statistical Association **108**(501), 278–287 (2013)

[19] Lu, C., Chang, K., Singh, P., Kalpathy-Cramer, J.: Three applications of conformal prediction for rating breast density in mammography (2022). https://doi.org/10.48550/ARXIV.2206.12008, `https://arxiv.org/abs/2206.12008`

[20] Lu, C., Lemay, A., Chang, K., Hoebel, K., Kalpathy-Cramer, J.: Fair conformal predictors for applications in medical imaging (2022)

[21] Lu, J., Pedemonte, S., Bizzo, B., Doyle, S., Andriole, K.P., Michalski, M.H., Gonzalez, R.G., Pomerantz, S.R.: Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. CoRR **abs/1807.10215** (2018)

[22] Michelmore, R., Wicker, M., Laurenti, L., Cardelli, L., Gal, Y., Kwiatkowska, M.: Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control. CoRR **abs/1909.09884** (2019)

[23] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F.A., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. arXiv preprint arXiv:2106.07998 (2021)

[24] Romano, Y., Patterson, E., Candès, E.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, vol. 32, pp. 3543–3553. NIPS (2019)

[25] Romano, Y., Sesia, M., Candès, E.: Classification with valid and adaptive coverage. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 3581–3591. Curran Associates, Inc. (2020)

[26] Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**, 223 – 234 (2019)

[27] Sesia, M., Romano, Y.: Conformal prediction using conditional histograms. Advances in Neural Information Processing Systems **34** (2021)

[28] Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(Mar), 371–421 (2008)

[29] Shashikumar, S., Wardi, G., Malhotra, A., Nemati, S.: Artificial intelligence sepsis prediction algorithm learns to say "I don't know". npj Digital Medicine **4** (12 2021)

[30] Vazquez, J., Facelli, J.: Conformal prediction in clinical medical sciences. Journal of Healthcare Informatics Research (01 2022)

[31] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York, NY, USA (2005)

[32] Vovk, V., Gammerman, A., Saunders, C.: Machine-learning applications of algorithmic randomness. In: International Conference on Machine Learning. pp. 444–453 (1999)

A Formal Theorem Statements and Proofs

Theorem 1 (Formal conformal coverage guarantee)**.** *Let* (X_1, Y_1)*,* (X_2, Y_2)*, ...,* (X_n, Y_n) *and* $(X_{\text{test}}, Y_{\text{test}})$ *be drawn independently and identically distributed from distribution* $\mathbb{P}$*, and let* $\mathcal{C}_\lambda$ *be any sequence of sets nested in* λ *such that* $\lim_{\lambda \to \infty} \mathcal{C}_\lambda = \mathcal{Y}$*. Finally, let* $\hat{\lambda} = \mathcal{A}(\mathcal{C}_\lambda,\ \alpha)$*. Then* $\mathcal{C}_{\hat{\lambda}}$ *satisfies* (3)*.*

Theorem 1. This is a standard result in the conformal literature [32]. A stripped down proof appears in [2]. The nested set version appears in [13]. □

The informal version of Theorem 1 in the main text is simply a corollary of the above when applied to the specific sequence of nested sets $\mathcal{T}_\lambda$.

Corollary 1. Pick $\lambda^{(1)} \leq \lambda^{(2)}$. It is clear that $\hat{f}(j|x) \geq \lambda^{(2)}$ implies $\hat{f}(j|x) \geq \lambda^{(1)}$. Therefore $\mathcal{T}_{\lambda^{(2)}}(x) \subseteq \mathcal{T}_{\lambda^{(1)}}(x)$. Applying Theorem 1 completes the proof. □

Corollary A.1. *In the setting of Theorem 1, let* $\tilde{\mathcal{T}}_\lambda$ *be the sequence of nested sets computed using Algorithm 1, and let* $\hat{\lambda} = \mathcal{A}\Big(\tilde{\mathcal{T}}_\lambda,\ \alpha\Big)$*. Then* $\tilde{\mathcal{T}}_{\hat{\lambda}}$ *satisfies 3.*

Corollary A.1. Pick $\lambda^{(1)} \leq \lambda^{(2)}$. Examining Algorithm 1, $q \leq \lambda^{(1)}$ implies $q \leq \lambda^{(2)}$. Therefore $\tilde{\mathcal{T}}_{\lambda^{(1)}}(x) \subseteq \tilde{\mathcal{T}}_{\lambda^{(2)}}(x)$. Applying Theorem 1 completes the proof. □