Fair Conformal Predictors for Applications in Medical Imaging

Charles Lu,[1] Andreanne Lemay,[2] Ken Chang,[3] Kathi Hoebel,[3] Jayashree Kalpathy-Cramer[1]
[1] Department of Radiology, Massachusetts General Hospital
clu@mgh.harvard.edu
[2] Polytechnique Montreal
[3] Massachusetts Institute of Technology

Abstract

Deep learning has the potential to augment many components of the clinical workflow, such as medical image interpretation. However, the translation of these black box algorithms into clinical practice has been marred by the relative lack of transparency compared to conventional machine learning methods, hindering in clinician trust in the systems for critical medical decision-making. Specifically, common deep learning approaches do not have intuitive ways of expressing uncertainty with respect to cases that might require further human review. Furthermore, the possibility of algorithmic bias has caused hesitancy regarding the use of developed algorithms in clinical settings. To these ends, we explore how conformal methods can complement deep learning models by providing both clinically intuitive way (by means of confidence prediction sets) of expressing model uncertainty as well as facilitating model transparency in clinical workflows. In this paper, we conduct a field survey with clinicians to assess clinical use-cases of conformal predictions. Next, we conduct experiments with a mammographic breast density and dermatology photography datasets to demonstrate the utility of conformal predictions in ”rule-in” and ”rule-out” disease scenarios. Further, we show that conformal predictors can be used to equalize coverage with respect to patient demographics such as race and skin tone. We find that a conformal predictions to be a promising framework with potential to increase clinical usability and transparency for better collaboration between deep learning algorithms and clinicians.

Introduction

The cost of healthcare continues to increase in several countries as populations age and lifespans increase. There has been substantial interest and investment recently in the fields of machine learning and computer vision, with the hope of increasing the efficiency and efficacy of clinical workflows. However, translating promising AI research into actual clinical use has been a difficult endeavour due to a general mistrust of these technologies by users, with increasing scepticism in feasibility of realizing the expansive promises to transform healthcare with AI technology(Strickland 2019). Evidently, relatively few clinical AI products have been cleared by the Federal and Drug Administration (FDA) in the United States and even fewer see wide deployment into clinical practice.

The challenge of designing clinical AI applications, not usually encountered or anticipated in traditional medical device software, has started to be more widely acknowledged by agencies such as the FDA and increasingly studied by the human-computer interaction (HCI) community (Jacobs et al. 2021; Xie et al. 2020). One such challenge of current deep learning approaches is the inability of identifying cases with high uncertainty that require further manual review. Further, lack of transparency in black box models has been cited as a barrier to developing AI tools that will be adopted by clinicians(Yang, Steinfeld, and Zimmerman 2019). Specifically, transparency has become a bigger concern in light of recent studies that have highlighted the high risk of algorithmic bias (Pierson et al. 2021b; Banerjee et al. 2021).

To address these challenges, we explore the use of a conformal prediction framework with the goal of increasing clinician trust in AI models. The impetus for exploring this framework comes from the observation that doctors routinely express uncertainty in the form of a set (i.e. differential diagnosis) while current machine learning uncertainty approaches express uncertainty as (somewhat arbitrary) numerical values (Hoebel et al. 2020). This difference makes current uncertainty methods less digestible for the user. Conformal methods, which output a set of predictions, better match the intuition of decision-making while providing statistical guarantees about the calibration of predictions that can better incorporated into safety-critical applications(Messoudi, Rousseau, and Destercke 2020). The formal coverage that conformal methods provide may also be potentially useful in mitigating algorithmic biases in order to avoid sustaining and exacerbating existing healthcare disparities.

As a first step in this work, we conduct field interviews with clinicians to understand how conformal methods can compliment clinical use-cases. Then, we evaluate our framework on two real world medical imaging datasets for mammographic breast density and dermatological photographs of diverse pathologies. We demonstrate how the conformal framework can be adapted for rule-in and rule-out disease scenarios. We also propose a conformal classifier that provides equalized coverage for known subgroups, such as race and skin tone, and compare with other common uncertainty quantification techniques.

Our results establish the feasibility of the conformal framework for classification tasks in medical imaging applications in breast density screening and skin lesion diagnosis that can be easily extended to other domains and tasks, such as regression and image segmentation. Importantly, we show the utility of this approach in presenting a more clinically intuitive way of presenting uncertainty as well as increasing the transparency of deep learning models.

Our contributions are the following:

1. Conduct field interviews with radiologist and clinicians to collect insight into clinical use-cases where conformal methods might be utilized
2. Detail how conformal methods could be incorporated for two clinically applicable use-cases: *rule-in* and *rule-out*.
3. Evaluate conformal predictors for coverage disparity in race and skin tone patient subpopulations in real-world mammography and dermatology datasets

Related work

Conformal predictions

The conformal framework is a general paradigm of generating prediction sets that does not place any assumptions on the particular model or data distribution. First introduced by Vovk, Gammerman, and Shafer (2005), conformal predictors can be calibrated in such a way to guarantee marginal coverage,

$$1 - \alpha \ \leq \ Pr(Y \ \in \ C(X)) \qquad (1)$$

where $C \subseteq \mathcal{P}(\{1, 2, \ldots, K\})$ is the prediction set, a subset of the power set of possible classes, Y, and $\alpha \in (0, 1)$ is a choosen confidence level. A conformal predictor is considered *valid* if the prediction set covers the true target, Y, with probability $1 - \alpha$ on the joint distribution, P_{XY}.

Any classifier that learns a score function, such as a convolutional neural network with softmax, $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, can be adapted in the conformal framework to output a prediction set, constructed by $\{y \in Y \mid S(X, Y) \leq \hat{q}\}$, where $\hat{q}$ is the quantile of scores which achieves the desired α coverage (Sadinle, Lei, and Wasserman 2019).

Typically a supervised classification model would only output a single score of the predicted class (e.g. *"basal cell carcinoma"*, 0.99) which may not actually correspond to a well-calibrated probability; however, a conformal model predicts a set that contains the true class with a statistically rigorous confidence level (e.g. {*"basal cell carcinoma", "squamous cell carcinoma", "seborrheic keratosis"*} with 95% coverage)

Interestingly, marginal coverage can be guaranteed for any arbitrary classifier over the joint probability of X and Y if $\hat{q}$ is chosen according to a calibration set of data, $\{X, Y\}_{i=1}^{n}$ drawn exchangeably.

$$1 - \alpha \ \leq \ Pr(Y_{n+1} \ \in \ C(X_{n+1})) \ \leq \ 1 - \alpha + \frac{1}{n+1} \qquad (2)$$

where, $C(X_{n+1}) = \{y \in S(X_{n+1}, y) > \hat{q} \mid y \in \mathcal{Y}\}$ and n the number of examples in the calibration set. The marginal bound follows from (Vovk, Gammerman, and Shafer 2005) and appears in several other works (Sadinle, Lei, and Wasserman 2019; Angelopoulos et al. 2020; Romano et al. 2020; Shafer and Vovk 2008).

Algorithm 1: Conformal Predictions

```
Input: Calibration set $\{(X, Y)\}_{i=1}^{n}$,
Test example $X_{n+1}$,
Confidence level $\alpha \in (0, 1)$,
Scoring function $S : X \times Y \to \mathbb{R}^{Y}$,
Quantile function $Q : \{\mathbb{R}\}_{i=1}^{n} \times [0, 1] \to \mathbb{R}$
Output: Prediction set $C(X_{n+1})$

for $i \in \{1, 2, \ldots, n\}$ do
    $s_i \leftarrow S(x_i, y_i)_{Y=y_i}$
end for
$\hat{q}_\alpha \leftarrow Q(\{s_i\}_{i=1}^{n}, \lceil (1-\alpha)(1+n) \rceil)$
return : $\{y : S(X_{n+1}, Y = y) > \hat{q}_\alpha \mid y \in Y\}$
```

Fairness metrics

A known issue in AI is the potential of algorithmic bias to learn to encode and exacerbate preexisting disparities between patient demographics in the healthcare system (Larrazabal et al. 2020; Pierson et al. 2021a; Seyyed-Kalantari et al. 2021). To evaluate potential and mitigate algorithmic bias, formal definitions of fairness have been developed to evaluate and mitigate fairness(Barocas, Hardt, and Narayanan 2019). Several common definitions of fairness measure statistical parity between mutually exclusive groups such as race or sex. Group fairness metrics for classification, $Y \in \{0, 1\}$, include *demographic parity*,

$$Pr(\hat{Y} = 1 \mid A = a) \ = \ Pr(\hat{Y} = 1 \mid A = b), \qquad (3)$$

which desires independence between the group attribute, $a, b \in A$, and the predicted response $\hat{Y}$, *equalized odds*

$$Pr(\hat{Y} = 1 \mid Y = y, A = a) \ = \ Pr(\hat{Y} = 1 \mid Y = y, A = b), \qquad (4)$$

for all classes, $y \in Y$, which desires conditional independence, $A \perp \hat{Y} \mid Y$, and *calibration parity*, which is achieved if the predicted scores, $S \in [0, 1]$, of all groups are perfectly well-calibrated

$$s = Pr(Y = y \mid S = s, A = a) = Pr(Y = y \mid S = s, A = b) \qquad (5)$$

for all score thresholds, $\forall s \in S$.

As shown in Kleinberg, Mullainathan, and Raghavan (2016), not all fairness metrics are mutually satisfiable, and many metrics assume the outcome for all groups is known and measurable, a condition that cannot be fulfilled in many clinical contexts. Techniques to mitigate different forms of bias assume access to the model during training or make explicit assumptions about the data distribution, which would likely also be infeasible in a clinical setting where models may be regulated and controlled by third-party medical device manufacturers(Zafar et al. 2017; Kusner et al. 2017).

Epistemic Uncertainty

Deep learning models are difficult to inspect for failure modes and known to be vulnerable to a number of attacks that can compromise safety and privacy of patients(Qayyum et al. 2021). To better design and integrate AI applications into existing clinical workflows and increase transparency of black box models, Bhatt et al. (2020) promotes uncertainty quantification to facilitate trust and mitigate bias in protected patient subgroups (Lu et al. 2021).

Deep neural networks are not regarded to be well-calibrated and temperature (Platt) scaling have been shown to help calibrate the softmax probability scores(Guo et al. 2017).

$$\text{softmax} = \frac{e^{z/\beta}}{\sum_i e^{z_i/\beta}} \tag{6}$$

where z is the logit output and β is a learned scalar that minimizes negative log-likelihood on the validation set.

As detailed in Hendrycks and Gimpel (2017), an extremely simple baseline for uncertainty, originally used to detect out-of-distribution examples and distribution shift, is the maximum softmax probability (MSP),

$$\text{MSP}(p) = 1 - \max(p). \tag{7}$$

As full Bayesian neural networks are difficult to scale, other techniques such as deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and stochastic weight averaging gradient (SWAG) (Maddox et al. 2019) also approximate the posterior by averaging models and fitting a Gaussian to multiple local minimas, respectively. Introduced by Gal and Ghahramani (2016), Monte Carlo dropout is a popular Bayesian approximation in neural networks that keeps dropout activated during inference time to sample from the posterior, $\{p_t\}_{t=1}^T = p_1, p_2, \ldots, p_t,$. Common quantitative measures of epistemic uncertainty include maximum softmax probability, predictive variance, and predictive entropy: **predictive variance** is defined as the average variance over T samples,

$$\text{Var}\left(\{p\}_{t=1}^T\right) = \frac{1}{C}\sum_{c=1}^{C} \text{Var}(p_c), \tag{8}$$

where $Var(p_c) = \frac{1}{T}\sum_{t=1}^{T}\left(p_{t|c} - \frac{1}{T}\sum_{t=1}^{T} p_{t|c}]\right)^2$, ans **predictive entropy** is the expected information over T samples,

$$\text{H}\left(\{p\}_{t=1}^T\right) = -\frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T} p_{t|c} \log \frac{1}{T}\sum_{t=1}^{T} p_{t|c}\right). \tag{9}$$

Field Interviews

To better study different clinical workflows and understand how radiologists think about uncertainty as it relates to AI, we conduct semi-structured interviews with four clinicians with varying levels of experience in interpreting medical images (see table **??**). The primary objective of these field interviews was to better understand the role of AI interaction in the clinical decision-making context in order to develop different prototypes of use-cases for conformal methods. We recruited participants from a large hospital system in a snowball sampling approach and obtained informed consent from all participants. Each interview lasted between 30 and 90 minutes.

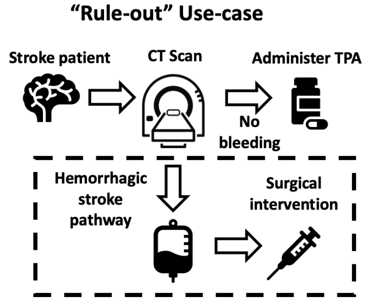

Figure 1: In patient suspected of stroke, a CT Head protocol is ordered to rule-out hemorrhagic strokes which is a contraindication for administering TPA medication for ischemic stroke and change to the alternative hemorrhagic stroke pathway shown in dashed box.

When asked about how AI predictions should be presented to the user, one clinician commented,

> Don’t overwhelm the radiologist with 83% or 87%. Is it good enough or is it not? [...] You just wanna say is this ”highly confident”, ”cautionary”, ”uncertain”, or ”non-diagnostic”? You have to set cutoffs but you don’t have to show those numbers to radiologists.

In addition to the concern of information overload, one clinician suggested that prediction outputs should be mapped to semantically meaningful concepts instead of raw probability scores. A distinction was made between two modes of clinical use-cases, with a radiologist commenting,

> I got to look at all the other stuff and decide whether it’s normal or just typical expected abnormalities for age and describe those, [...] eventually there will be AI algorithms that are setup to evaluate all those things and sometimes your sensitivity and your thresholds for what you expect are different if you’re just checking if everything looks normal versus if it was a study done to specifically evaluate that.

An imaging exam is often ordered by a referring physician to **rule-out** critical but low-probability findings as part of a standard decision tree in a clinical pathway. For example, a computed tomography (CT) study is typically done in patients with suspected stroke to rule-out hemorrhagic bleeding in the brain that would prevent the administration of tissue plasminogen activator (TPA) to break down blood clots, which is the usual treatment pathway for ischemic stroke.

An AI algorithm to detect hemorrhagic bleeding should have high sensitivity (true positive rate) as administering TPA in patients with hemorrhagic stroke would be disastrous by worsening the bleeding.

Another possible AI use-case would be to **rule-in** serious conditions and screen out patients with milder conditions in

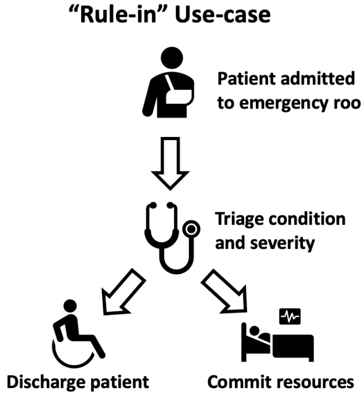
"Rule-in" Use-case
Patient admitted to emergency room
Triage condition and severity
Discharge patient
Commit resources

Figure 2: Workflow of an AI algorithm triage in the emergency room to expedite patients with no severe conditions.

order to better allocate limited resources. One clincian gave an example of triage in an emergency department:

> They want to discharge patients from the ER as quickly as possible – healthy patients that don't need to be there – bed space is limited. Sometimes they're waiting to get an answer from radiology. So one of the workflows is the algorithm identifies normal cases – no disease – but you want it to identify patients that were definitely normal with very high confidence. You want the high specificity. If the doctor is overseeing 20 patients, and the algorithms said these 3 have very low chance of having any abnormality. Those ones, you may feel comfortable discharging with just your look at them instead of waiting on the radiologist.

On the issue of fairness and bias, one clinician saw the need for platforms to audit AI algorithms and dashboards that could display different types of metadata for clinically relevant demographics, such as age and race, to check for bias. Monitoring and reviewing model predictions would be needed to determine necessary adjustments to thresholds to recalibrate models to change disease prevalence and populations over time.

Fair Conformal Predictors

Because conformal predictors provide a statistically valid prediction set, one notion of fair predictors is assuring that each subgroup has the same average coverage (Romano et al. 2020). Equal coverage has several advantages over statistical parity metrics in clinical AI applications. First, the base rates of disease prevalence and risk factors differ between relevant patient demographics, such as race, gender, and age, so that attempting to mitigate disparity via parity metrics may actually negatively impact patient outcomes. Second, as highlighted in our field interviews, different clinical use cases prioritize different types of errors, which have different downstream consequences in the care pathway. Outputting a prediction set allows greater flexibility in intended use-cases, such as triage and screening, and gives the clinician more insight by considering alternative predicted conditions similar to a differential diagnosis.

Algorithm 2: Group adaptive prediction sets (GAPS)

Input: Calibration set $\{(X, A, Y)\}_{i=1}^{n}$,
Test example $X_{n+1,A=a'}$,
Confidence level $\alpha \in (0, 1)$,
Scoring function $S : X \times A \times Y \to \mathbb{R}^{Y}$,
Quantile function $Q : \{\mathbb{R}\}_{i=1}^{n} \times [0, 1] \to \mathbb{R}$
Sorting function SORT : $\mathcal{S} \to \mathcal{S}$
Output: Confidence prediction set $C(X_{n+1} \mid A = a')$

```
for $a \in A$ do
    for $i \in \{1, 2, \dots, n\}$ do
        $s_{a,i} \leftarrow \sum_{j=1}^{y_j=y_i}$ SORT$(S(x_i, y_i \mid A = a)$
    end for
    $\hat{q}_{a,\alpha} \leftarrow Q(\{s_{a,i}\}_{i=1}^{n}, \lceil(1-\alpha)(1+n)\rceil)$
end for
return : $\{y : S(X_{n+1}, A = a', Y = y) > \hat{q}_{a,\alpha} \mid y \in Y\}$
```

We now describe an conformal algorithm to achieve equalized marginal coverage at some confidence $1 - \alpha$ on a subgroup A in a multi-class setting over the joint distribution $P(X, A, Y)$. Note that the conformal framework extends to other prediction tasks such as quantile regression, multilabel classification, and image segmentation (Romano, Patterson, and Candès 2019; Cauchois, Gupta, and Duchi 2020; Bates et al. 2021). Assuming the subgroup membership is known at inference time (a mild assumption in the clinical setting as patient history and medical records would usually be available), we adapt the conformal algorithm by calibrating each subgroup on their respective calibration set and apply a separate quantile, $\hat{q}_a$ at inference time, preserving coverage as proved in Romano et al. (2020). We modify the approach to conformal classification from Romano, Sesia, and Candes (2020) for *group adaptive prediction sets* (GAPS).

Let $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$ be a set of three tuples containing a training example $X_i \in \mathrm{R}^d$, group attribute $A_i \in \{1, 2, \dots K\}$, and target label $Y_i \in \{0, 1, \dots, C\}$. Assuming all samples are drawn interchangeably from an arbitrary distribution $\mathcal{D}_{XAY}$, then marginal coverage is assured for each subgroup,

$$1-\alpha \le Pr(Y_{n+1} \in C(X_{n+1}, A_{n+1} \mid A_{n+1} = a) \le 1-\alpha+\frac{1}{n+1} \tag{10}$$

for all $a \in A,\ \alpha \in (0, 1)$.

We also propose a fairness metric given a conformal predictor, $f : \mathcal{X} \to \mathcal{C}$ and α confidence level, then *coverage disparity* can be defined as the pairwise difference between subgroup coverage:

$$\frac{1}{\mid A \mid} \sum_{a,b \in \binom{A}{2}} \left|\text{Cov}(f(X \mid A = a) - \text{Cov}(f(X \mid A = b)\right|, \tag{11}$$

where Cov is the coverage of a predictor and $\binom{A}{2}$ is all combinations of unique pairs of subgroups.

Table 1: Percentage of breast density according to BI-RADS scale by patient race

RACE	Fatty	Scattered	Hetero.	Dense	Count
WHITE	10.9%	43.5%	39.2%	6.4%	86881
BLACK	14.8%	48.5%	33%	3.7%	15047
ASIAN	7.1%	30.5%	49.7%	12.7%	3978
LATINO	13.5%	50.4%	32.2%	3.9%	1514
OTHER	12.1%	44.6%	35.9%	7.4%	810

Table 2: Percentage of skin conditions in three dermatological categories by Fitzpatrick skin tone scale

SCALE	Benign	Malignant	Non-neoplastic	Count
1	15%	15.4%	69.6%	2947
2	14%	15.4%	70.6%	4808
3	14.4%	13.8%	72%	3308
4	13.2%	10.8%	76%	2781
5	10.4%	9.6%	80%	1533
6	6.9%	9.6%	83.5%	635
MISSING	13.1%	18.2%	68.7%	565

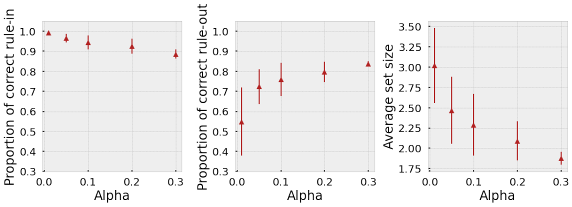

Figure 3: Ruling-in and ruling-out the most dense breasts out of four grades of breast density using GAPS for conformal predictions.

Experiments

We utilize the multi-institutional Digital Mammographic Imaging Screening Trial (DMIST) dataset (Pisano et al. 2005), utilizing 12-bit monochrome 1, 12-bit monochrome 2, and 14-bit monochrome 1 images as previously described (Chang et al. 2020). Our patient cohort consisted of 108,230 digital screening images from 21,759 patients. Accurate assessment of breast density, as measured by relative amounts of fibroglandular tissue, is important for the interpretation of mammography and is commonly rated according to the BI-RADS criteria into one of four categories: entirely fatty, scattered, heterogeneously dense, and extremely dense (Liberman and Menell 2002). The ”extremely dense” category is of particular clinical importance because these patients may have masked cancers and thus may benefit from supplemental imaging (Bakker et al. 2019). As such, we consider the ”extremely dense” class as the critical class and the other three classes as non-critical. We consider reported patient race as a subgroup to analyze: ”White”, ”Black or African American”, ”Latino or Hispanic”, ”Asian”, and ”Other / Unknown”.

The Fitzpatrick17k dataset aggregates 16,577 photography images collected from two dermatology atlases. Hierarchical labels of 114 different skin conditions aggregated into three categories (benign lesions, malignant lesions, and non-neoplastic lesions) are supplemented with the Fitzpatrick skin type, an ordinal 6-point scale that is a proxy for the amount of melanin pigment in the skin (Groh et al. 2021).Skin tone in image classification systems has been demonstrated to impact performance and underperform on darker skin tones in several facial recognition systems(Buolamwini and Gebru 2018). In addition to biases in data collection and model development, prevalence of various skin conditions, such as melanoma, is known to differ based on skin tone. To investigate differences between skin color within the context of our two use-cases, we evaluate conformal methods, treating the 11 malignant skin conditions as critical classes and the other conditions as non-critical.

We compare several different approaches of predictions sets: *naive* (Algorithm 1), *adaptive prediction sets* (APS) (Romano, Sesia, and Candes 2020), *regularized adaptive prediction sets* (RAPS) (Angelopoulos et al. 2020) with our *group adaptive prediction sets* (GAPS) described in 2. For our deep learning image classifier, we use the ResNet50 and Wide-ResNet(50-101) architectures (Xie et al. 2017; Zagoruyko and Komodakis 2016) in our experiments and average over 5 runs with different random seeds using Monte Ccarlo dropout of 0.1 and $T = 30$ to calculate epistemic uncertainty metrics. The models were trained with a cross-entropy loss for 100 epochs with an early stopping after 20 epochs, learning rate of 0.0001, and batch size of 16. As a naive baseline, we utilized Platt scaling with maximum softmax probability. We implemented all experiments using the PyTorch framework (Paszke et al. 2019) and trained all models on a Nvidia A100 GPU. Related code with relavant hyperparameters used in our experiments are available at **Gihub repo**.

Results

On the DMIST dataset, we find high performance (near or above 90% accuracy) for ruling in patients with extremely dense breasts at confidence levels between 99% and 70% (Figure 3). A favorable trade-off in sensitivity, specificity, and cardinality appears to be between $\alpha = 0.2$ and $\alpha = 0.3$. Comparing different conformal techniques (Naive, APS, RAPS, GAPS) on patient race subgroups, we find greater coverage than the expected theoretical bound for at confidence levels beyond $\alpha = 0.01$ across the all four conformal methods (depicted in Figure 4). Coverage of Naive and RAPS drop more steeply as α increases while APS and GAPS maintain higher coverage levels, which do not fall below 60% for any level of confidence. Interestingly, no significant difference in coverage rate appears between subgroups for any of the four conformal techniques at most α thresholds.

Further investigating the distribution of ruling-in and ruling-out use-cases using the equalized coverage GAPS

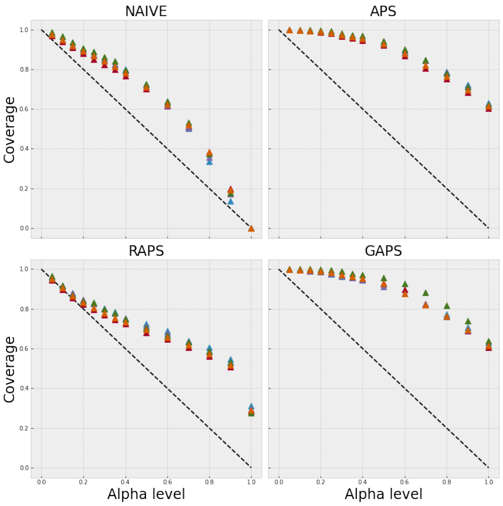

Figure 4: Subgroup coverage at different α thresholds on DMIST dataset. Colors denotes patient race. Dashed line denotes the theoretical marginal coverage.

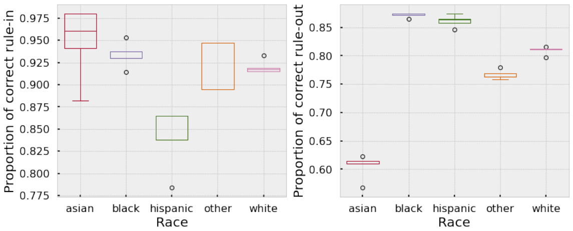

Figure 5: Boxplots of ruling-in and ruling-out extremely dense breast density by patient race (conformal predictions with GAPS at $\alpha = 0.1$).

method at 90% confidence level in Figure 9, we find relative difference in accuracy and variance between races. The majority subgroup (white patients) has lower variance in both use-cases and more similar performance between ruling-in and ruling-out. Moving on to conformal's relationship with other uncertainty metrics in Figure 6, we find the cardinality (prediction set size) to be positively correlated with maximum softmax probability and predictive entropy, but not strongly correlated with predictive variance.

On the Fitzpatrick17k dataset, we find much lower variance estimates than in the previous dataset in Figure 7. High confidence levels ($\alpha = 0.05$) have sharp drop in ruling-out malignant skin lesions and large prediction set sizes indicating higher uncertainty in detecting one of the 11 possible malignant conditions out of 114 different skin conditions. In contrast breast density and race, we find more linear coverage for skin lesion classification with different skin tones in all four conformal techniques (Figure 8). Again, APS and GAPS provide higher coverage at lower α than Naive and RAPS. Notably, for the Naive conformal method, the theoretical coverage guarantee is not satisfied on the lighter skin tones.

As shown in Figure 10, we again observe that prediction set size is positively correlated with maximum softmax probability and predictive entropy but not with predictive variance. We find no significant difference between critical and non-critical cases, however, we notice heteroscedasticity in both maximum softmax and predictive variance as cardinality increases.

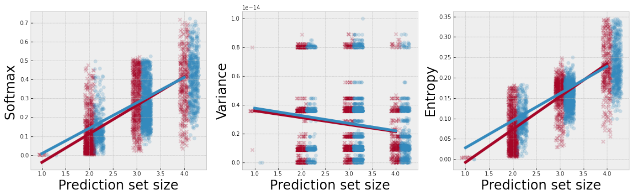

Figure 6: Correlation of 2000 randomly sampled points on the DMIST mammography dataset (red points are extremely breast density and blue points are other densities).

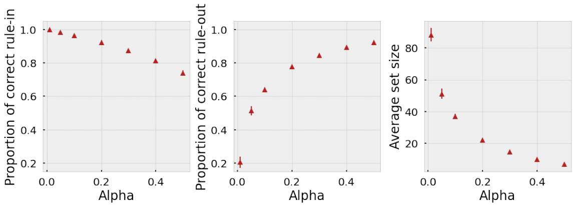

Figure 7: Ruling-in and ruling-out 11 types of malignant lesions out of 114 total skin conditions using GAPS for conformal predictions.

Table 3 compares disparity as measured by our metric of coverage disparity defined in Equation 11. For breast density rating on the DMIST dataset, both APS and GAPS show lower disparity in the race subgroup, however no significant difference in coverage disparity was found in skin tone between different conformal methods on the skin lesion classification.

Discussion

In this study, we explore the potential of conformal predictions for healthcare applications. From our field interviews, we gain insight into how conformal predictions better map onto clinical intuitions about decision-making than maximum likelihood scores or arbitrary values of epistemic uncertainty. Furthermore, we identify two general and clinically useful categories of clinical use-cases for conformal predictors: rule-out and rule-in. The results of our survey suggested that a conformal framework would allow for greater transparency to the user in identifying possible clinical mimics in a differential diagnosis, thus improving the interaction experience between AI models and human users.

As a proof of concept, we validate the feasibility of conformal predictors for two clinical applications within the medical domains of radiology and dermatology with datasets containing vastly different subgroup distributions (highly imbalanced for race and balanced for skin color). Our modified conformal method (GAPS) performs well for both use-cases of ruling-in and ruling-out critical conditions

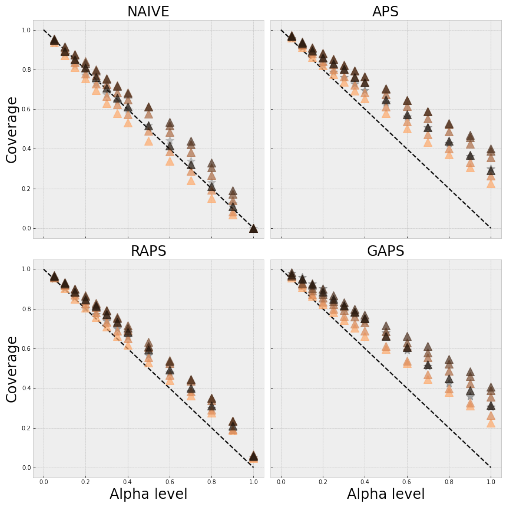

Figure 8: Subgroup coverage at different α thresholds on Fitzpatrick dataset. Skin colors are Fitzpatrick groups. Gray star is the subgroup with missing skin tone.

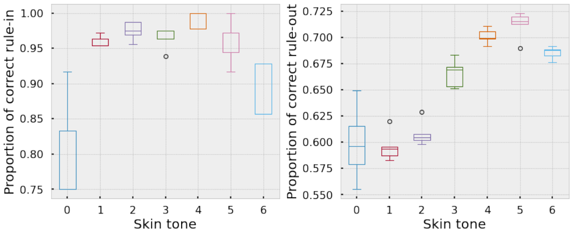

Figure 9: Boxplots of ruling-in and ruling-out malignant skin lesions by Fitzpatrick skin tone (0: no Fitzpatrick skin tone label available) (conformal predictions with GAPS at $\alpha = 0.1$).

such as breast cancer risk and malignant skin diseases. Importantly, we argue that conformal fairness is more suitable than other fairness metrics, which assume direct access to treatment outcomes or model – an unrealistic assumption in a real clinical setting in which AI tools would be deployed by external medical device manufacturers and the original training scheme and dataset would be unknown or unavailable. Equalized subgroup coverage is more semantically meaningful in clinical contexts, such as ruling-in or ruling-out critical disease conditions. Furthermore, conformal notions of uncertainty such as prediction set size and confidence level may be an attractive alternative to other quantification metrics of epistemic uncertainty, which do not provide any formal guarantees of confidence or calibration, and may contradict each other (see predictive variance and entropy in above experiments).

Overall, our results suggest that the fair conformal predictors have the potential to increase clinician trust in AI models by incorporating meaningful assurances of confidence between clinically relevant sub-populations such as race and skin tone. Although our proof-of-concept experiments focused on classification, conformal inference is easily adapted to the other tasks such as regression and segmentation (Romano et al. 2020; Bates et al. 2021). Additionally, the conformal prediction framework can encompass any AI model that outputs a score function and many machine learning paradigms, such as decision trees and and deep learning, and does no assumptions on the data distribution.

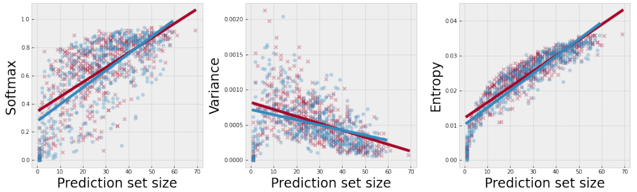

Figure 10: Correlation between 500 randomly sampled points on the Fitzpatrick dermatology dataset (red points patients with malignant skin lesions and blue points are patients with non-malignant skin conditions).

Table 3: Coverage disparity of different conformal prediction methods at difference confidence levels. Lower values are better.

$1-\alpha$	**95%**	**85%**	**75%**
Coverage Disparity in Race in Breast Density Rating			
NAIVE	0.008 ± 0.005	0.013 ± 0.008	0.016 ± 0.009
APS	0.001 ± 0.001	0.003 ± 0.003	0.006 ± 0.005
RAPS	0.007 ± 0.004	0.011 ± 0.006	0.014 ± 0.008
GAPS	0.001 ± 0.001	0.004 ± 0.003	0.009 ± 0.006
Coverage Disparity in Skin Tone in Lesion Classification			
NAIVE	0.009 ± 0.007	0.030 ± 0.018	0.046 ± 0.027
APS	0.007 ± 0.004	0.021 ± 0.013	0.037 ± 0.022
RAPS	0.008 ± 0.005	0.023 ± 0.014	0.035 ± 0.021
GAPS	0.011 ± 0.007	0.031 ± 0.019	0.036 ± 0.022

Some limitations of conformal predictors still need to be addressed in future work such as choosing cutoffs that correspond to clinically meaningful guidelines. Further characterizing adaptiveness of conformal prediction sets will be useful in determining how to best operationalize these methods for clinical practice. Lastly, more user studies are needed to more rigorously evaluate clinical feasibility and usability in the clinical workflow of the radiologist, which will be dependent of the clinical use case and entail prospective testing on data from multiple sites with large heterogeneous populations.

We believe clinical AI in safety-critical domains like medicine require calibrated estimates of uncertainty to justify clinical decision-making and conformal inference can be used to better design and develop more robust and safer AI systems. We hope this work promotes further work and efforts into their applicability for clinical applications to impact healthcare and improve patient outcomes.

References

Angelopoulos, A. N.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty Sets for Image Classifiers using Conformal Prediction.

Bakker, M. F.; de Lange, S. V.; Pijnappel, R. M.; Mann, R. M.; Peeters, P. H.; Monninkhof, E. M.; Emaus, M. J.; Loo, C. E.; Bisschops, R. H.; Lobbes, M. B.; et al. 2019. Supplemental MRI screening for women with extremely dense breast tissue. *New England Journal of Medicine*, 381(22): 2091–2102.

Banerjee, I.; Bhimireddy, A. R.; Burns, J. L.; Celi, L. A.; Chen, L.-C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; Kuo, P.-C.; et al. 2021. Reading Race: AI Recognises Patient's Racial Identity In Medical Images. *arXiv preprint arXiv:2107.10356*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bates, S.; Angelopoulos, A. N.; Lei, L.; Malik, J.; and Jordan, M. I. 2021. Distribution Free, Risk Controlling Prediction Sets.

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G. G.; Krishnan, R.; Stanley, J.; Tickoo, O.; Nachman, L.; Chunara, R.; Srikumar, M.; Weller, A.; and Xiang, A. 2020. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Cauchois, M.; Gupta, S.; and Duchi, J. 2020. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. arXiv:2004.10181.

Chang, K.; Beers, A. L.; Brink, L.; Patel, J. B.; Singh, P.; Arun, N. T.; Hoebel, K. V.; Gaw, N.; Shah, M.; Pisano, E. D.; et al. 2020. Multi-institutional assessment and crowdsourcing evaluation of deep learning for automated classification of breast density. *Journal of the American College of Radiology*, 17(12): 1653–1662.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, New York, USA: PMLR.

Groh, M.; Harris, C.; Soenksen, L.; Lau, F.; Han, R.; Kim, A.; Koochek, A.; and Badri, O. 2021. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. *arXiv preprint arXiv:2104.09957*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330. PMLR.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *Proceedings of International Conference on Learning Representations*.

Hoebel, K.; Andrearczyk, V.; Beers, A.; Patel, J.; Chang, K.; Depeursinge, A.; Müller, H.; and Kalpathy-Cramer, J. 2020. An exploration of uncertainty information for segmentation quality assessment. In *Medical Imaging 2020: Image Processing*, volume 11313, 113131K. International Society for Optics and Photonics.

Jacobs, M.; He, J.; Pradier, M. F.; Lam, B.; Ahn, A. C.; McCoy, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *CHI*, 659:1–659:14.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR*, abs/1609.05807.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

Liberman, L.; and Menell, J. H. 2002. Breast imaging reporting and data system (BI-RADS). *Radiologic clinics of North America*, 40: 409–430.

Lu, C.; Lemay, A.; Hoebel, K.; and Kalpathy-Cramer, J. 2021. Evaluating subgroup disparity using epistemic uncertainty in mammography. arXiv:2107.02716.

Maddox, W. J.; Izmailov, P.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Messoudi, S.; Rousseau, S.; and Destercke, S. 2020. Deep Conformal Prediction for Robust Models. In Lesot, M.-J.; Vieira, S.; Reformat, M. Z.; Carvalho, J. P.; Wilbik, A.; Bouchon-Meunier, B.; and Yager, R. R., eds., *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 528–540. Cham: Springer International Publishing. ISBN 978-3-030-50146-4.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Pierson, E.; Cutler, D.; Leskovec, J.; Mullainathan, S.; and Obermeyer, Z. 2021a. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27: 136–140.

Pierson, E.; Cutler, D. M.; Leskovec, J.; Mullainathan, S.; and Obermeyer, Z. 2021b. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1): 136–140.

Pisano, E. D.; Gatsonis, C.; Hendrick, E.; Yaffe, M.; Baum, J. K.; Acharyya, S.; Conant, E. F.; Fajardo, L. L.; Bassett, L.; D'Orsi, C.; Jong, R.; and Rebner, M. 2005. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *New England Journal of Medicine*, 353(17): 1773–1783.

Qayyum, A.; Qadir, J.; Bilal, M.; and Al-Fuqaha, A. 2021. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering*, 14: 156–180.

Romano, Y.; Barber, R. F.; Sabatti, C.; and Candès, E. 2020. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2). Https://hdsr.mitpress.mit.edu/pub/qedrwcz3.

Romano, Y.; Patterson, E.; and Candès, E. J. 2019. Conformalized Quantile Regression. arXiv:1905.03222.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with Valid and Adaptive Coverage. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3581–3591. Curran Associates, Inc.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least Ambiguous Set-Valued Classifiers With Bounded Error Levels. *Journal of the American Statistical Association*, 114(525): 223–234.

Seyyed-Kalantari, L.; Liu, G.; McDermott, M. B. A.; and Ghassemi, M. 2021. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 26: 232–243.

Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12): 371–421.

Strickland, E. 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, 56: 24–31.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*.

Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. 5987–5995.

Xie, Y.; Chen, M.; Kao, D.; Gao, G.; and Chen, X. A. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Yang, Q.; Steinfeld, A.; and Zimmerman, J. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment and Disparate Impact. *Proceedings of the 26th International Conference on World Wide Web*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. *CoRR*, abs/1605.07146.