Fair Conformal Predictors for Applications in Medical Imaging

Charles Lu,[1, †] **Andréanne Lemay,**[2] **Ken Chang,**[3, 4] **Katharina Höbel,**[3, 4] **Jayashree Kalpathy-Cramer**[1, 4]

[1] Athinoula A. Martinos Center for Biomedical Imaging
[2] Polytechnique Montréal
[3] Massachusetts Institute of Technology
[4] Harvard Medical School
[†] clu@mgh.harvard.edu

Abstract

Deep learning has the potential to automate many clinically useful tasks in medical imaging. However translation of deep learning into clinical practice has been hindered by issues such as lack of the transparency and interpretability in these "black box" algorithms compared to traditional statistical methods. Specifically, many clinical deep learning models lack rigorous and robust techniques for conveying certainty (or lack thereof) in their predictions – ultimately limiting their appeal for extensive use in medical decision-making. Furthermore, numerous demonstrations of algorithmic bias have increased hesitancy towards deployment of deep learning for clinical applications. To this end, we explore how conformal predictions can complement existing deep learning approaches by providing an intuitive way of expressing uncertainty while facilitating greater transparency to clinical users. In this paper, we conduct field interviews with radiologists to assess possible use-cases for conformal predictors. Using insights gathered from these interviews, we devise two clinical use-cases and empirically evaluate several methods of conformal predictions on a dermatology photography dataset for skin lesion classification. We show how to modify conformal predictions to be more adaptive to subgroup differences in patient skin tones through equalized coverage. Finally, we compare conformal prediction against measures of epistemic uncertainty.

Introduction

As the cost of healthcare continues to rise in many countries around the world, massive investment has been poured into medical AI development in the hopes of leveraging these technologies to lower costs and improve efficiency (Zhang et al. 2021; Topol 2019). However translation into clinical deployment has proved extremely difficult – exemplified by several highly publicized projects ultimately failing to be widely used in clinical practice (Strickland 2019).

Medical AI tools can pose unique design considerations that significantly differ from more traditional medical device software (Lu et al. 2021a). Regulatory agencies such as the U.S. Food and Drug Administration (FDA) have begun to consider these differences with more scrutiny (Food and Administration 2021). Research communities such as human-computer interaction also increasingly explore the

unique challenges of designing machine learning for healthcare applications (Jacobs et al. 2021; Xie et al. 2020; Beede et al. 2020).

One challenging obstacle for clinical machine learning is identifying cases in which the model has low confidence – a scenario that might necessitate further review of those cases. Furthermore, the lack of transparency in machine learning models has also been cited as a major barrier to developing AI tools that are adopted by clinicians (Yang, Steinfeld, and Zimmerman 2019; Begoli, Bhattacharya, and Kusnezov 2019). A survey by Allen et al. (2021) showed that even though many radiologists believed that AI tools added value to their clinical practice, they would not trust AI for autonomous clinical use.

Deep learning models, in particular, are notoriously difficult to interrogate and known to be vulnerable to a number of attacks that can compromise the safety and privacy of patients (Arun et al. 2021; Finlayson et al. 2019). Also, many recent studies highlight the risks of algorithmic biases in machine learning for applications such as skin lesion classification and facial recognition systems (Bissoto, Valle, and Avila 2020; Buolamwini and Gebru 2018; Pierson et al. 2021b; Banerjee et al. 2021).

To address these challenges, we explore the use of conformal predictions methods for medical imaging applications. We argue that conformal predictions may better correspond to clinical decision-making intuition and can be easily incorporated into existing models while providing meaningful statistical guarantees (Messoudi, Rousseau, and Destercke 2020). For example, doctors routinely express uncertainty in the form of comparative sets to arrive at a diagnosis i.e. a *differential diagnosis*. Additionally, conformal prediction sets naturally convey a measure of uncertainty by the number of items contained in their set. Several works promote uncertainty quantification to facilitate greater trust in medical black-box models and to detect bias in protected patient demographics (Bhatt et al. 2021; Kompa, Snoek, and Beam 2021; Lu et al. 2021b).

In this study, we demonstrate the potential of conformal predictors to offer a more clinically intuitive representation of model uncertainty. Based on interviews with clinicians, we characterize two conformal use-cases (*rule-in* and *rule-out*) and show how conformal predictors can be adapted to yield equalized coverage at a subgroup level. We empiri-

cally evaluate our group conformal predictors on a dermatology dataset for skin lesion classification using Fitzpatrick skin type as our group attribute. Finally, we compare conformal uncertainty against epistemic uncertainty to show how group-calibrated conformal predictors better represent relevant subgroup differences such as disease prevalence of malignant skin conditions.

Related Work

Fairness in Machine Learning

A major challenge for the deployment of machine learning in healthcare is the potential to encode and and reinforce inequalities among protected patient demographics (Larrazabal et al. 2020; Pierson et al. 2021a). Attempts have been made to formally conceptualize "fairness" statistically and definitions of fairness have been proposed to measure disparity between demographic groups such as race and gender (Barocas, Hardt, and Narayanan 2019).

Examples of group fairness metrics for binary classification, $Y \in \{0, 1\}$, include *demographic parity*:

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

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

$$P(\hat{Y} = 1 \mid Y = y, A = a) \ = \ P(\hat{Y} = 1 \mid Y = y, A = b), \quad (2)$$

for $y \in \{0, 1\}$ and desires conditional independence, $A \perp \hat{Y} \mid Y$, and *calibration parity*, which is achieved only if the prediction scores, $S \in [0, 1]$, for each subgroup are perfectly calibrated

$$s = P(Y = y \mid S = s, A = a) = P(Y = y \mid S = s, A = b)$$
$$(3)$$

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

As shown in Kleinberg, Mullainathan, and Raghavan (2017), not all fairness metrics are mutually satisfiable, and many metrics assume the observed outcome for all subgroups is known and measurable (an assumption that may not be practical in many clinical situations). Furthermore, many bias mitigation strategies often assume access to the model's training procedure or make explicit distributional assumptions (Zafar et al. 2017; Kusner et al. 2017). For regulated medical device software, these assumptions may not be realistic as medical device manufacturers would not permit such access or modifications to their proprietary, commercial models.

Uncertainty for Deep Learning

Epistemic uncertainty concerns the probabilistic estimation of model parameters. Several approaches to uncertainty have been developed for deep neural networks, such as Bayesian deep learning, which encompasses different approaches to approximating full Bayesian neural networks (Gal 2016). Techniques such as Monte-Carlo dropout modify the architecture or training procedure of the model to learn a posterior distribution over the parameters (Gal and Ghahramani 2016). For example, an ensemble of deep models can be trained with adversarial examples and then averaged to estimate the posterior distribution (Lakshminarayanan, Pritzel, and Blundell 2017).

Regardless of which technique is used, the posterior distribution can be randomly sampled to estimate different measures of epistemic uncertainty, such as *predictive entropy*, which is the average entropy over T samples and K classes,

$$-\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t=1}^{T}p_{t|k}\log\frac{1}{T}\sum_{t=1}^{T}p_{t|k}\right). \qquad (4)$$

An alternative approach to uncertainty involves calibrating model predictions such that the confidence of a prediction matches the actual probability that that prediction would be correct for all confidence levels. Deep neural networks are generally regarded to be poorly calibrated. This has motivated many post-hoc methods, such as Platt Scaling, to help better calibrate the outputs of the softmax function (Guo et al. 2017):

$$\mathrm{Platt}(z) = \frac{e^{z/\beta}}{\sum_i e^{z_i/\beta}}, \qquad (5)$$

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

Conformal Predictions

One limitation of epistemic uncertainty is its lack of any standard or intuitive meaning to aid users in their decision making. While post-hoc calibration techniques do have a statistical interpretation, they do not usually provide any distribution-free guarantees of reliability or miscalibration. Additionally, common approaches often require access to the model (e.g. training with a modified loss function) or distributional assumptions (e.g. sampling from a Bernoulli distribution), which may not be feasible in third-party medical device software. Conformal prediction methods do not have these limitations.

First introduced by Vovk, Gammerman, and Shafer (2005), conformal prediction is a method of distribution-free uncertainty quantification, which is model agnostic and provides formal results for *marginal coverage*, defined as the average probability that the true class will be contained in the prediction set. Formally, this coverage guarantees states:

$$1 - \alpha \ \leq \ P\left(y_n \in \{y \in \mathcal{Y} \mid S(x_n)_y > \hat{q}\}\right), \qquad (6)$$

where α is some predetermined miscoverage level, y_n is the true class, x_n is the unseen test example, $S(x_n)_y$ is the yth index of the output score, and $\hat{q}$ is the estimated score quantile needed to achieve proper coverage at the desired level. The score quantile is estimated on a held-out set (assumed to be IID with the test set) to ensure marginal coverage for any classifier over the joint distribution, $D_{X,Y}$.

Any classifier that learns a score function, such as a deep neural network, can be adapted to output conformal predictions. For example, instead of outputting the predicted class with the highest score, e.g. ("basal cell carcinoma", 0.51), a conformal prediction with a miscoverage level of 10%, $\alpha = 0.1$, might produce a set of predictions, e.g. {("basal cell carcinoma", 0.51), ("squamous cell

participant	role	experience (years)
P1	neuroradiologist	20
P2	radiologist	15
P3	neurologist	10
P4	clinical fellow	5

Table 1: Role and years of experience with medical imaging of the four clinicians in our field interviews.

carcinoma", 0.27), ("seborrheic keratosis", 0.13)}, which will contain the true class on average 90% of the time.

These prediction sets naturally lends themselves to be utilized as a form of uncertainty and are well-suited for medical decision making (Kompa, Snoek, and Beam 2021; Vazquez and Facelli 2022). Although we focus on multi-class classification, the conformal framework is general and has been extended to other prediction tasks such as regression and multi-label classification (Romano, Patterson, and Candes 2019; Cauchois, Gupta, and Duchi 2021).

Field Interviews

To better investigate how conformal predictors could be utilized in clinical workflows, we conduct semi-structured interviews with four clinicians from a large hospital system (their role and level of experience is shown in Table 1). By asking participants to characterize their clinical workflow, we observed a distinction being made in the motivating reason for the imaging study, with one radiologist (P1) commenting:

> I got to [...] 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.

When asked how predictions of AI medical device software should be presented within the radiologist's user interface, P2 responded:

> 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 wanting AI outputs to be more semantically meaningful, clinicians were also concerned about the reliability and confidence of the model's predictions. P1 and P3 even suggested that low confidence predictions should not be presented to the user at all depending on the clinical context.

For example, a computed tomography (CT) imaging exam is typically performed on patients with suspected stroke to rule-out hemorrhagic bleeding, a condition that would prevent administration of tissue plasminogen activator (TPA) (Figure 1). Thus, an application to detect hemorrhagic
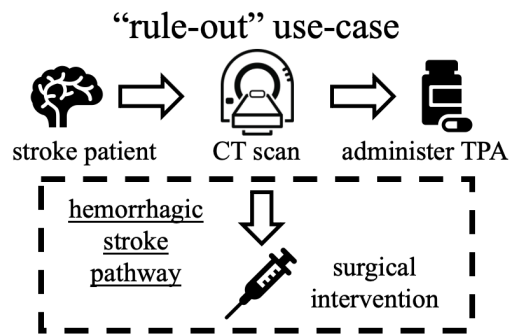

"rule-out" use-case

Figure 1: A diagram of typical stroke care pathway, in which a CT exam study is ordered to "rule-out" hemorrhagic stroke, a contraindication for clot-breaking medication (TPA), which is the usual treatment for ischemic stroke.
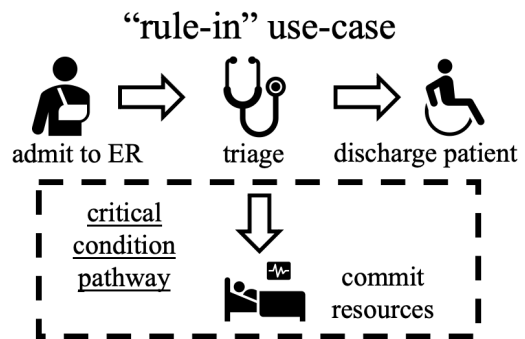

"rule-in" use-case

Figure 2: A depiction of a triage workflow in the emergency room to expedite patients without severe conditions and to "rule-in" patients with one or more critical conditions.

bleeds must prioritize sensitivity (true positive rate), even if the prediction confidence level had to be lowered, because giving TPA to a patient with a hemorrhagic stroke would worsen the bleeding. This use-case of confidently "ruling-out" critical conditions seems to us an ideal setting for conformal prediction.

From our interviews, another clinical example was given: the initial triage of patients 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.

To better allocate limited resources, patient with milder conditions may be discharged to outpatient care whereas patients more likely to have one or more severe conditions may

need more immediate or invasive intervention (Figure 2). Therefore, we formulate another use-case for conformal predictors to "rule-in" critical conditions. This "rule-in" use-case seems ideal for conformal predictors since the set of possible critical conditions can be very large and naturally lends itself to thinking in terms of prediction sets and possible interactions between present conditions (some conditions are of low concern when present separately but may be of higher concern if present together).

Lastly, on the issue of fairness and bias, P4 saw the need for auditing platforms and monitoring dashboards to display metadata for clinically relevant demographics, such as age and race, to check for bias. Continual review of model predictions would be necessary to adjust decision thresholds in order to recalibrate models to changing disease prevalence over time.

Methods

As conformal methods provide distribution-free coverage guarantees regardless of the number of data examples or how the machine learning model was trained, they can be easily extended to provide *equalized coverage* with the group attribute is known. Introduced by Romano et al. (2020) for the regression setting, equalized subgroup coverage can adapt better to differences between subgroups than aggregate conformal methods.

While we acknowledge that any definition of fairness will be reductive and problematic with respect to some criteria, conformal fairness has several advantages for measuring disparity in medical AI tools. First, the prevalence of many disease naturally differ between some patient demographics, such as race and age, due to various genetic, social, and behavioral risk factors. Attempting to encapsulate these differences solely with a statistical parity measure without the necessary clinical context would be misguided and may mislead clinical decision-making. Instead, outputting a set of likely predictions allows greater flexibility in clinical use-cases, such as triage and screening, by providing the clinical user more intuition to discern what should be considered normal or abnormal, on a patient level as well as a group level. Practically, distribution-free uncertainty quantification techniques like conformal prediction can be applied to calibrate and monitor differences between patient cohorts in arbitrary medical AI devices for real-time clinical workflows.

We now describe a conformal algorithm to achieve equalized coverage for some coverage level, α, on an attribute group A. Assume that group membership is known at inference time, we can then estimate the empirical quantile, $\hat{q}_a$, for each subgroup separately on their respective held-out set and, at inference time, apply that subgroup's score quantile.

For example, we can modify the conformal approach adaptive prediction sets (APS) from Romano, Sesia, and Candes (2020) for the group setting: 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 class $Y_i \in \{0, 1, \dots, K\}$, where K is the number of classes. The scoring function s sorts the classes according to their softmax score in descending order for each example and outputs the cumulative sum until the softmax score of the true class is

---

Algorithm 1: Group adaptive prediction sets (GAPS)

Input: Held-out calibration set $\{(X, A, Y)\}^n \in \mathcal{X}^n \times \mathcal{A}^n \times \mathcal{Y}^n$,
Miscoverage level $\alpha \in (0, 1)$,
Test example $(x_{n+1}, a_{n+1}) \in \mathcal{X} \times \mathcal{A}$,
Number of classes $k := |\mathcal{Y}|$,
Prediction function $f : X \to \mathbb{R}^k$,
Scoring function $s : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$,
Sorting function SORT : $\mathbb{R}^k \to \mathbb{R}^k$,
Quantile function $q : \{\mathbb{R}\}^m \times [0, 1] \to \mathbb{R}$

1: **for** $a \in A$ **do**
2:     $s_a \leftarrow \{ (f(X_i), Y_i) : i \in \{1, 2, \dots, n\} \mid A_i = a\}$
3:     $q_a \leftarrow q(s_a, \frac{\lceil (1-\alpha) \cdot (n+1) \rceil}{n})$
4: **end for**
5: **for** $a \in A$ **do**
6:     **if** $a = a_{n+1}$ **then**
7:         $y_{n+1} \leftarrow f(x_{n+1})$
8:         $y' \leftarrow \text{SORT}(y_{n+1})$
9:         $c_{n+1} \leftarrow \{y'_j : j \in \{1, 2, \dots, k\} \mid \sum_{j=1}^k y'_j < q_a\}$
10:     **end if**
11: **end for**
12: **return** c_{n+1}

Output: Conformal prediction set c_{n+1}

---

reached. Assuming all samples are drawn interchangeably from a potentially unknown distribution, then marginal coverage can be assured for each subgroup,

$$1 - \alpha \le P(Y_{n+1} \in C(X_{n+1} \mid A = a) \le 1 - \alpha + \frac{1}{n+1} \quad (7)$$

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

Experiments

We compare four methods: a non-conformal baseline (Naive), adaptive prediction sets (APS) (Romano, Sesia, and Candes 2020), regularized adaptive prediction sets (RAPS) (Angelopoulos et al. 2021), group adaptive prediction sets (GAPS), and group regularized adaptive prediction sets (GRAPS). In our experiments, we investigate the following questions:

1. How do different methods compare in *rule-in* and *rule-out* use cases for malignant skin lesion classification?

2. How do group conformal predictors compare with aggregate conformal predictors conditioned on skin type as the group attribute?

3. What is the relationship between prediction set size and epistemic uncertainty?

Skin Lesion Dataset

For our experiments, we use the Fitzpatrick17k dataset, which aggregates $16{,}577$ photography images from two dermatology atlases (Groh et al. 2021). This dataset contains a hierarchical labeling scheme of 114 different skin conditions, which are further aggregated into three dermatological categories: non-neoplastic lesions (12,080), benign lesions (2,234), and malignant lesions (2,253). Additionally, each
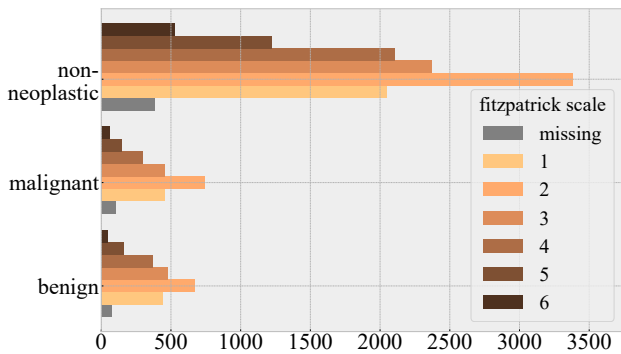
Figure 3: Distribution of skin conditions by Fitzpatrick skin type and categorization of the 114 different lesions into one of three broad categories: non-neoplastic, malignant, or benign.

skin type	non-neoplastic	benign	malignant
1	69.6%	15.0%	15.4%
2	70.5%	14.0%	15.5%
3	71.8%	14.4%	13.8%
4	75.9%	13.2%	10.8%
5	80.0%	10.4%	9.6%
6	83.4%	7.0%	9.6%
missing	68.9%	13.0%	18.1%
total	72.8%	13.5%	13.7%

Table 2: Percentage of skin conditions in each of the three broad dermatological categories for different Fitzpatrick skin type subgroups.

image is labeled with a Fitzpatrick skin type – an ordinal 6-point scale that estimates the amount of melanin pigment in the skin. The disease distribution for each skin type is shown in Figure 3 and Table 2.

As the prevalence of skin conditions such as melanoma is known to differ between demographics such as race and ethnicity, we treat the Fitzpatrick skin type as the group attribute in our experiments. We also treat the 11 malignant skin conditions as "critical" conditions and the other 103 conditions as "non-critical" for the purposes of our two conformal prediction use-cases. For the "rule-in" scenario, only critical cases with one of the 11 malignant skin condition were considered, with a prediction was considered correct if the true class was contained in the prediction set. For the "rule-out" scenario, only non-critical cases were considered, with a prediction was considered correct if none of the 11 malignant conditions were contained in the prediction set.

Evaluation Metrics

We consider our methods with two evaluation metrics: *marginal coverage* and *set size*. Given some miscoverage parameter, α, marginal coverage is defined as the number of times the true class is contained in the prediction set, averaged over all examples.

Set size is the average number of elements in the predic-

tion set. To quantify difference between subgroups, we also define *coverage disparity* to be the average pairwise difference in marginal coverage

$$\frac{1}{|A|} \sum_{a,b \in \binom{A}{2}} \left| \text{coverage}_{A=a} - \text{coverage}_{A=b} \right|, \qquad (8)$$

where $\binom{A}{2}$ is all pairwise combinations of subgroups.

We define *set size disparity* in a similar manner

$$\frac{1}{|A|} \sum_{a,b \in \binom{A}{2}} \left| \text{set size}_{A=a} - \text{set size}_{A=b} \right|. \qquad (9)$$

Implementation Details

We use a ResNet-18 (Xie et al. 2016), pretrained on ImageNet (Deng et al. 2009), as our deep learning image classifier and average across five different initializations. We use Monte Carlo dropout of 0.1 and $T = 30$ to estimate maximum softmax probability (Hendrycks and Gimpel 2017) and predictive entropy (Equation 4) with softmax scores calibrated using Platt scaling (Equation 5). For our non-conformal baseline, we form prediction sets by sorting softmax scores and taking classes until the $1 - \alpha$ threshold. All models were trained with cross-entropy for 100 epochs, early stopping of 20 epochs, a learning rate of 0.0001, and a batch size of 16. All experiments were implemented using the PyTorch framework (Paszke et al. 2019) and trained on a Nvidia A100 GPU. Code and data to reproduce our experiments is made available here: https://github.com/clu5/AAAI22.

Results

For our two use-cases, we find all four methods of prediction sets to perform similarly at $\alpha = 0.1$ in terms of accuracy and set size at a subgroup level (Figure 4). Group conformal methods (GAPS and GRAPS) in performing slightly better at ruling-out malignant skin conditions and worse at ruling-out malignant skin conditions than other methods for skin types 1 and 2 (the lightest skin tones).

Next, we compare performance between Fitzpatrick skin types in terms of marginal coverage and set size between aggregate conformal methods (APS and RAPS) and their group variants (GAPS and GRAPS) in Figure 5. We observe lower subgroup coverage disparity as well as higher set size disparity with GAPS and GRAPS compared to APS and RAPS (Table 3).

Finally, we consider the relationship between set size and epistemic uncertainty, measured by maximum softmax probability and predictive entropy, and find a strong correlation between set size and epistemic uncertainty. Looking at Figure 6, we can see that the group conformal methods show increased separation between Fitzpatrick subgroups – particularly for skin type 1 and skin type 6 – compared to the naive baseline and aggregate conformal methods. This separation is quantified in Table 4, which shows that GAPS and GRAPS have lower Spearman correlation than Naive, APS, and RAPS across a range of α.
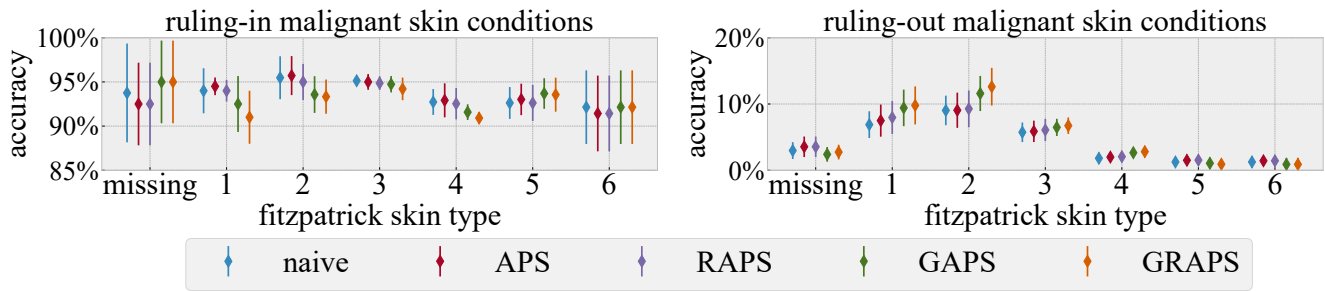
Figure 4: Subgroup accuracy of different prediction set methods at $\alpha = 0.1$ for ruling-in and ruling-out use-cases.
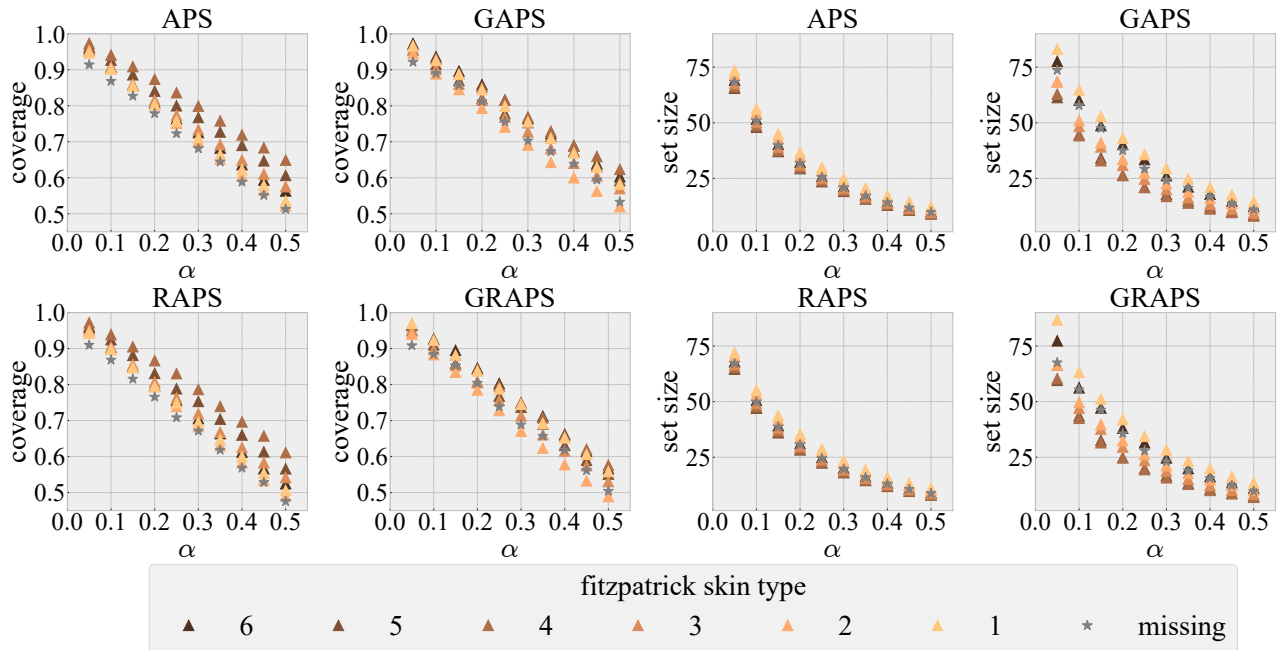

Figure 5: Comparing coverage and set size of group conformal methods (GAPS and GRAPS) and aggregate conformal methods (APS and RAPS) at different α values for skin lesion classification; colors denote Fitzpatrick skin types; gray star represents the missing skin type.

Discussion

In this paper, we explored might be conformal predictors integrated into medical imaging workflows. Our interviews suggested several ways in which conformal predictions could allow for greater transparency in medical AI, such as the identification of possible clinical mimics in a manner similar to differential diagnoses. We also formulated two general use-cases for conformal predictions for clinical decision-making.

To better calibrate conformal predictors to subgroup differences, we modified two methods of conformal predictions (APS and RAPS) to guarantee equalized coverage for known demographic attributes and performed experiments on a heterogeneous skin lesion dataset with Fitzpatrick skin type to evaluate group conformal predictors. Results from our experiments demonstrated that group conformal predictors have lower coverage disparity than aggregate conformal

methods and show greater variability in set size at a subgroup level, reflecting underlying differences between different skin types. Specifically, skin types 1 and 6 have larger prediction set sizes, indicating more uncertainty. These differences may be due to the fact that lesions of patients with lighter skin tones have a higher rate of malignancy (Table 2) while patients with the skin type 6 are underrepresented in the dataset, comprising only about 4% of the total dataset.

Further comparing group conformal uncertainty against aggregate conformal uncertainty, we observed that group conformal methods are less correlated with epistemic uncertainty measures, thus indicating more adaptiveness to subgroup differences across task difficulty levels as measured by maximum softmax probability and predictive entropy. This observation agrees with visual inspection of the scatter plots in Figure 6 that show increased separation for both malignant and non-malignant skin conditions for the lightest and darkest skin tones. Therefore, we conclude that group
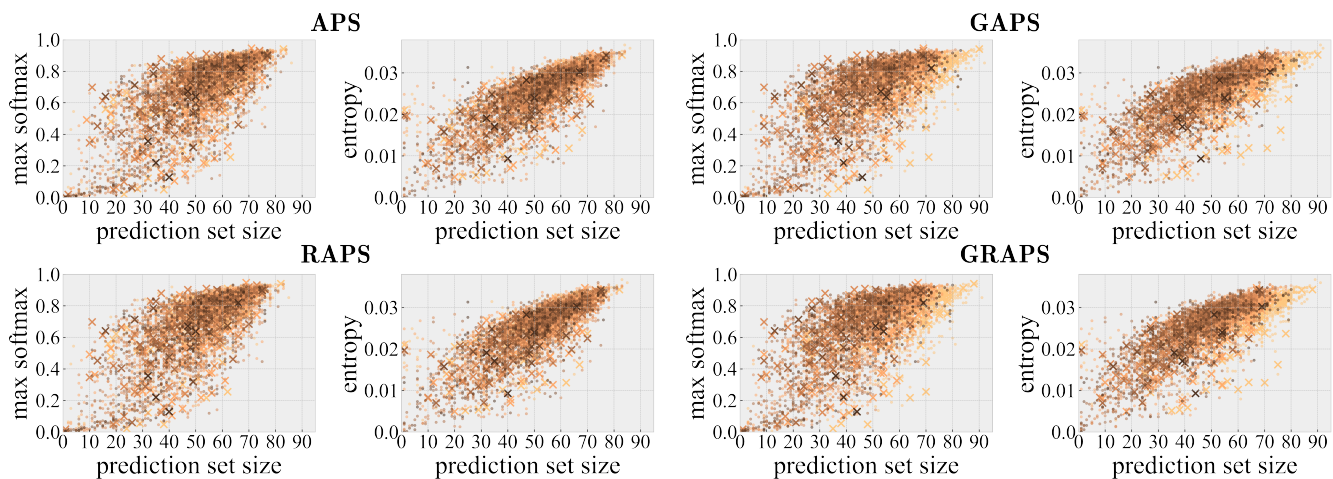
Figure 6: Relationship between conformal uncertainty and two epistemic uncertainty measures at $\alpha = 0.1$; colors correspond to different Fitzpatrick skin types; $\times$ denotes a malignant skin condition and $\circ$ denotes a non-malignant skin condition.

	method	0.1	0.2	0.3	0.4	0.5
coverage	NAIVE	0.026	0.040	0.059	0.070	0.080
	APS	0.026	0.035	0.048	0.056	0.059
	RAPS	0.026	0.037	0.049	0.054	0.057
	GAPS	0.024	0.029	0.037	0.038	0.045
	GRAPS	0.022	0.028	0.037	0.035	0.039
set size	NAIVE	3.9	2.2	1.2	0.6	0.2
	APS	3.5	3.1	2.4	1.9	1.4
	RAPS	3.6	3.0	2.4	1.8	1.3
	GAPS	9.7	8.1	5.6	4.2	2.8
	GRAPS	9.1	8.0	5.5	4.2	2.7

Table 3: Subgroup disparity in coverage and set size between Fitzpatrick skin types at five different levels of α.

	method	0.1	0.2	0.3	0.4	0.5
softmax	NAIVE	0.71	0.75	0.78	0.75	0.55
	APS	0.69	0.72	0.75	0.78	0.82
	RAPS	0.69	0.73	0.75	0.79	0.82
	GAPS	0.66	0.70	0.74	0.76	0.80
	GRAPS	0.66	0.70	0.74	0.76	0.80
entropy	NAIVE	0.82	0.84	0.80	0.66	0.34
	APS	0.81	0.83	0.85	0.87	0.87
	RAPS	0.81	0.84	0.85	0.87	0.87
	GAPS	0.77	0.81	0.84	0.84	0.86
	GRAPS	0.77	0.81	0.84	0.84	0.86

Table 4: Spearman correlation between set size and epistemic uncertainty (maximum softmax probability and predictive entropy) at five different values of α.

conformal methods can better describe subgroup uncertainty than regular conformal methods when subgroups come from different data distributions.

Conclusion

Fair conformal predictors have the potential to increase clinician trust in AI models by providing meaningful notions of uncertainty across clinically relevant sub-populations. These distribution-free methods complement existing deep learning models and require no modifications to existing training procedures. Based on our experiments, we find group conformal predictors to be a promising and generally applicable approach to increasing clinical usability and trustworthiness in medical AI. We hope this work promotes further work into group conformal predictors for clinical applications in healthcare.

Acknowledgements

References

Allen, B.; Agarwal, S.; Coombs, L. P.; Dreyer, K.; and Wald, C. 2021. 2020 ACR Data Science Institute Artificial Intelligence Survey. *Journal of the American College of Radiology*.

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

Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; Hoebel, K.; Gupta, S.; Patel, J.; Gidwani, M.; Adebayo, J.; Li, M. D.; and Kalpathy-Cramer, J. 2021. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3(6): e200267.

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. Accessed: 2021-04-11.

Beede, E.; Baylor, E.; Hersch, F.; Iurchenko, A.; Wilcox, L.; Ruamviboonsuk, P.; and Vardoulakis, L. M. 2020. *A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy*, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

Begoli, E.; Bhattacharya, T.; and Kusnezov, D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1: 20–23.

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.; and Xiang, A. 2021. *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*, 401–413. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Bissoto, A.; Valle, E.; and Avila, S. 2020. Debiasing Skin Lesion Datasets and Models? Not So Fast. In *ISIC Skin Image Anaylsis Workshop, 2020 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

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. C. 2021. Knowing what You Know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81): 1–42.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Finlayson, S. G.; Bowers, J. D.; Ito, J.; Zittrain, J. L.; Beam, A. L.; and Kohane, I. S. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289.

Food, U.; and Administration, D. 2021. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan.

Gal, Y. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

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. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1820–1828.

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*.

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.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., ed., *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 43:1–43:23. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3.

Kompa, B.; Snoek, J.; and Beam, A. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4.

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.

Lu, C.; Chang, K.; Singh, P.; Pomerantz, S. R.; Doyle, S.; Kakarmath, S.; Bridge, C. P.; and Kalpathy-Cramer, J. 2021a. Deploying clinical machine learning? Consider the following.. *CoRR*, abs/2109.06919.

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

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.

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 Candes, E. 2019. Conformalized Quantile Regression. 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.

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.

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

Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25: 44–56.

Vazquez, J.; and Facelli, J. 2022. Conformal Prediction in Clinical Medical Sciences. *Journal of Healthcare Informatics Research*.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*, chapter 2, 17–51. Springer.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2016. Aggregated Residual Transformations for Deep Neural Networks. *CoRR*, abs/1611.05431.

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*.

Zhang, D.; Mishra, S.; Brynjolfsson, E.; Etchemendy, J.; Ganguli, D.; Grosz, B.; Lyons, T.; Manyika, J.; Niebles, J. C.; Sellitto, M.; Shoham, Y.; Clark, J.; and Perrault, R. 2021. The AI Index 2021 Annual Report.