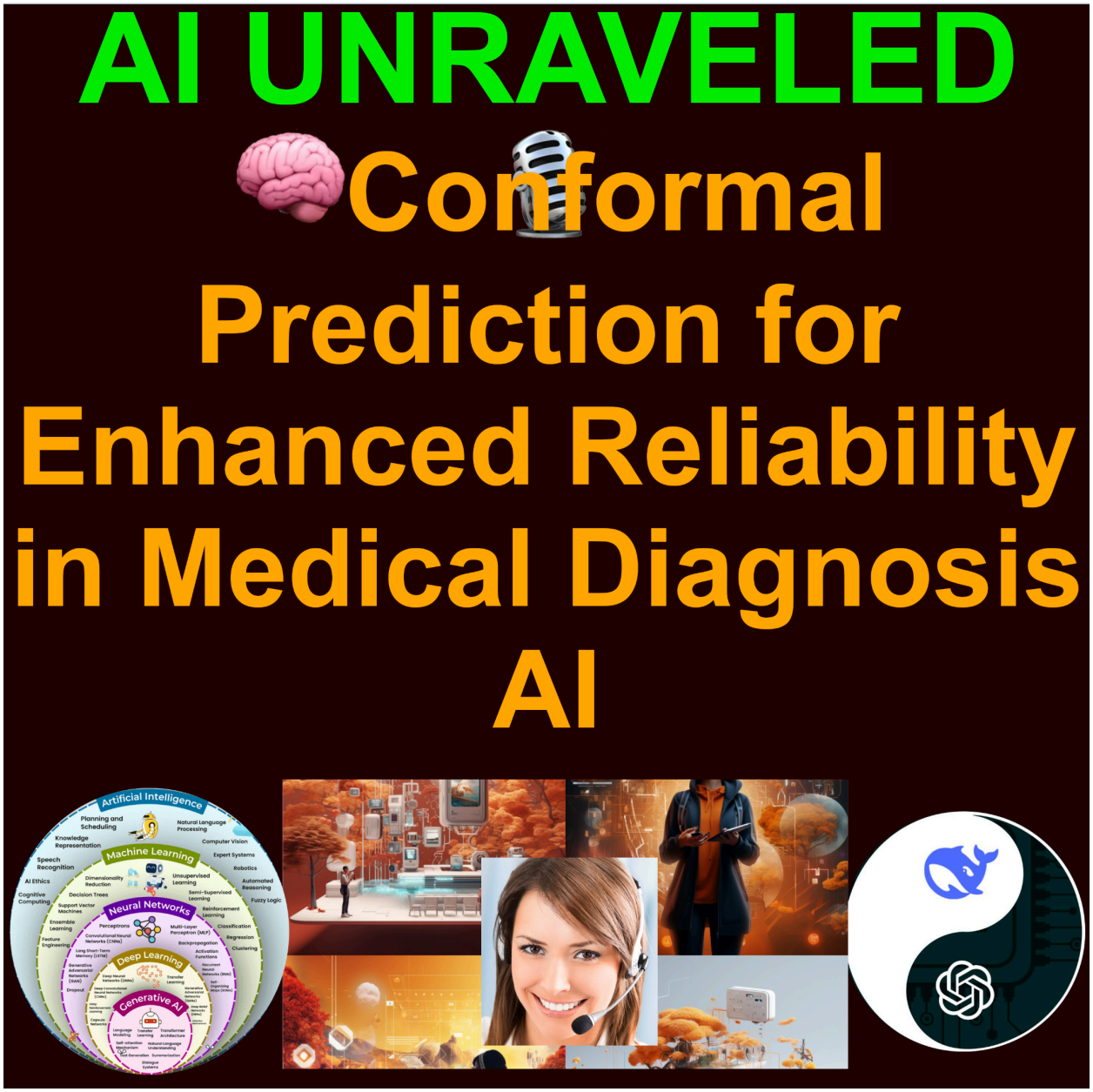
🧠Conformal Prediction for Enhanced Reliability in Medical Diagnosis AI



1. Introduction: The Imperative for Reliable AI in Medical Diagnosis

The Promise and Peril of AI in Healthcare

Artificial intelligence (AI) and machine learning (ML) hold immense promise for transforming healthcare, particularly in the realm of medical diagnosis. AI systems

demonstrate potential in analyzing complex medical data, including medical images, electronic health records (EHRs), genomic sequences, and health monitoring signals, often identifying subtle patterns beyond human perception.[1] Applications range from early disease detection and diagnosis to personalized treatment recommendations and advancements in biomedical research.[1] However, the deployment of AI in high-stakes clinical environments is fraught with challenges. The inherent unreliability of predictive models poses significant risks, as inaccurate predictions can lead to misdiagnosis, inappropriate treatment, and adverse patient outcomes.[1] The success of AI models has often been measured primarily by predictive accuracy, but in critical applications like medicine, accuracy alone is insufficient.[4] The potential for life-threatening consequences, such as those associated with delays in diagnosing severe bacterial infections which cause millions of deaths globally [5], underscores the critical need for AI systems that are not just powerful, but fundamentally reliable and trustworthy.

The Uncertainty Challenge

A core issue stems from the fact that AI models produce predictions, which are inherently uncertain [User Query]. This uncertainty arises from multiple sources. **Aleatoric uncertainty** reflects the inherent randomness or noise in the data itself, which cannot be reduced even with more data.[11] **Epistemic uncertainty**, on the other hand, arises from limitations in the model or insufficient training data; this type of uncertainty can potentially be reduced with better models or more data.[11] Effectively quantifying and communicating this uncertainty is paramount for building clinician trust, ensuring patient safety, and facilitating the responsible adoption of AI in clinical practice.[4] Without robust uncertainty quantification (UQ), AI models can produce overconfident incorrect predictions, hindering their clinical utility and potentially causing harm.[7]

The growing integration of AI into healthcare reflects a significant shift. Initially, the focus was predominantly on maximizing predictive accuracy.[4] However, limitations in model reliability and resulting lack of clinician trust have impeded widespread adoption, particularly in critical diagnostic pathways.[1] This has spurred a move towards demanding not just high performance, but *provable reliability*. This shift is driven by the recognition that for high-stakes decisions, heuristic confidence scores are inadequate. There is a need for methods that provide mathematically rigorous guarantees about prediction reliability, aligning with the increasing focus of regulatory bodies like the U.S. Food and Drug Administration (FDA) on the safety, effectiveness, and continuous monitoring of AI-based medical devices.[3]

Introduction to Uncertainty Quantification (UQ)

Uncertainty quantification (UQ) is the field dedicated to characterizing and managing the uncertainty associated with computational and experimental models.[4] Several approaches exist for UQ in machine learning. Bayesian methods, such as Bayesian Neural Networks, attempt to learn distributions over model parameters to capture uncertainty.[4] Approximations like Monte Carlo (MC) Dropout simulate Bayesian uncertainty by applying dropout during inference.[11] Ensemble methods train multiple models and use the variance in their predictions as an uncertainty measure.[12] While valuable, these methods often rely on specific distributional assumptions or provide guarantees that are asymptotic or approximate.

Introducing Conformal Prediction (CP)

Within the UQ landscape, Conformal Prediction (CP), also referred to as Conformal Classification in the context of classification tasks, has emerged as a powerful and distinct framework.[4] Its core promise is compelling: instead of providing a single point prediction, CP generates a **prediction set** (for classification) or **prediction interval** (for regression) that is guaranteed to contain the true, unknown outcome with a user-specified probability (e.g., 95%).[4] Crucially, this coverage guarantee is **distribution-free** and holds with **finite samples**, requiring only the mild assumption of data exchangeability.[4] A further significant advantage is CP's **model-agnostic** nature; it can be applied as a wrapper around virtually any pre-trained machine learning model without modification or retraining, making it highly versatile and easy to integrate.[4] CP's ability to provide these rigorous, mathematically sound confidence bounds directly addresses the critical need for demonstrable safety and trustworthiness in medical AI.

Roadmap

This report provides a comprehensive examination of conformal prediction applied to medical diagnosis. It begins by detailing the foundational concepts and mathematical framework of CP. It then explores recent advancements, particularly the use of Test-Time Augmentation to improve CP's efficiency. Subsequently, it delves into more advanced CP methods designed to offer stronger or more tailored guarantees. A comparative analysis situates CP within the broader landscape of UQ techniques. The report then surveys diverse applications of CP across various medical domains, followed by a discussion of the challenges and pathways for clinical integration and decision support. Finally, it addresses the crucial ethical and regulatory dimensions before concluding with future research directions.

2. Foundations of Conformal Prediction

Core Concept: From Point Predictions to Prediction Sets

The fundamental departure of conformal prediction from traditional predictive modeling lies in its output. Instead of generating a single "best guess" prediction (e.g., diagnosing a condition as lung cancer), CP produces a *set* of possible outcomes.[4] For a classification task like medical diagnosis based on an image X, the output is not a single label y, but a prediction set C(X) containing one or more plausible diagnoses, such as C(X)={pneumonia, heart failure}. For a regression task, like predicting a biomarker level, the output is a prediction interval [L(X),U(X)]. This set-based output inherently acknowledges and quantifies the model's uncertainty. In medical contexts where different conditions can exhibit similar features (e.g., on imaging), providing a set of possibilities is often more clinically relevant and safer than forcing a single, potentially incorrect, diagnosis [User Query].

Mathematical Framework

The validity of CP rests on a well-defined mathematical framework, primarily developed by Vovk, Gammerman, Shafer, and others.[6]

Exchangeability Assumption: The cornerstone assumption underpinning CP's guarantees is **exchangeability**.[4] A sequence of data points $(X1,Y1),...,(Xn,Yn),(Xn+1,Yn+1)$ is exchangeable if their joint probability distribution $P((X1,Y1),...,(Xn+1,Yn+1))$ remains unchanged under any permutation of the indices $1,...,n+1$. This is a weaker condition than assuming the data are independent and identically distributed (i.i.d.), although i.i.d. data is always exchangeable.[4] Exchangeability implies that the order of observations provides no statistical information. It is this property that ensures that the rank (or related p-value) of a new data point's "strangeness" relative to past data is uniformly distributed, forming the basis for calibration.[4]

However, the reliance on exchangeability can be a significant limitation in dynamic clinical settings. Clinical data streams, such as EHR data evolving over time, or imaging data acquired using changing protocols or patient populations, may violate this assumption.[4] Temporal dependencies, distribution shifts between the calibration data and future test data, or systematic differences between patient subgroups can all break exchangeability.[7] Such violations compromise the theoretical coverage guarantee, potentially leading to misleadingly narrow prediction sets and a false sense of security.[4] This necessitates careful validation of the exchangeability assumption in practice or the use of specialized CP methods designed to handle non-exchangeable

data or distribution shifts.[24]

Nonconformity Scores (NCS): At the heart of CP is the **nonconformity score** (NCS), denoted s(x,y) or αi.[4] This score quantifies how "atypical," "strange," or "nonconforming" a specific data point (x,y) is relative to a set of other data points, according to some underlying model or heuristic. The score function, also called a nonconformity measure, is chosen by the user. Higher scores typically indicate greater nonconformity (less typical).

Examples include:

- For classification, using a model's predicted probability p^(y∣x) for a potential class y: s(x,y)=1−p^(y∣x).[11] A higher score means the model assigned lower probability to that class.
- For regression, using the absolute residual from a model's prediction f^(x): s(x,y)=∣y−f^(x)∣.[11] A larger residual indicates greater nonconformity.

The choice of NCS is critical for the *efficiency* (informativeness, i.e., the size) of the resulting prediction sets, although it does not affect the *validity* (coverage guarantee).[27] A well-chosen NCS, one that accurately reflects the model's uncertainty for different inputs and potential outputs, will lead to smaller, more useful prediction sets. Conversely, a poorly chosen or uninformative NCS can result in excessively large sets, even if the coverage guarantee technically holds, thereby limiting the practical value of the CP output.[24] Much research focuses on designing effective NCS tailored to specific problems and models.[11]

Calibration and Quantiles/P-values: CP uses a set of **calibration data**, denoted Dcal={(Xi,Yi)}i=1ncal, to determine a threshold for the nonconformity scores. This calibration set must be exchangeable with the future test data points for the guarantee to hold. In split/inductive CP (discussed below), Dcal is separate from the data used to train the underlying model.[4]

The nonconformity scores are computed for all points in the calibration set: αi=s(Xi,Yi) for i=1,...,ncal. Given a desired significance level (or maximum error rate) α∈(0,1), CP calculates the (1−α) empirical quantile of these calibration scores. Specifically, the threshold q^ is set to the ⌈(1−α)(ncal+1)⌉/(ncal+1) quantile of the scores {α1,...,αncal}.[4]

Alternatively, one can think in terms of p-values.25 For a new test point Xtest and a hypothesized label y, calculate its nonconformity score αtest=s(Xtest,y). The p-value for this hypothesis is the fraction of calibration scores that are greater than or equal to αtest:
p(y)=ncal+1∣{i∈{1,...,ncal}:αi≥αtest}∣+1
The +1 terms account for the test point itself in the ranking.

Prediction Set Construction: The prediction set C(Xtest) for a new input Xtest includes all possible labels $y \in Y$ (or values in regression) whose nonconformity score s(Xtest,y) is less than or equal to the threshold $\hat{q}$, or equivalently, whose p-value p(y) is greater than the significance level α.11

$$ C(X_{test}) = \{ y \in \mathcal{Y} \mid s(X_{test}, y) \leq \hat{q} \} \equiv \{ y \in \mathcal{Y} \mid p(y) > \alpha \} $$

Intuitively, the set includes all potential outcomes that are "conforming enough" compared to the calibration data, based on the chosen error tolerance α.

The Marginal Coverage Guarantee: The central theoretical result of CP is the marginal coverage guarantee.4 For any new data point (X_{n+1},Y_{n+1}) drawn from the same exchangeable distribution as the calibration data, the probability that the true label Y_{n+1} is contained within the constructed prediction set $C(X_{n+1})$ is at least $1-\alpha$:

$$P(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha$$

This guarantee holds for any finite sample size ncal, is independent of the underlying data distribution P, and holds regardless of the chosen nonconformity score function or the underlying prediction model used to derive it.4 The term "marginal" signifies that this guarantee applies on average across all possible test points drawn from the distribution P. It does not guarantee that for a specific test point $X_{n+1}=x$, the conditional probability $P(Y_{n+1} \in C(x) \mid X_{n+1}=x)$ is exactly $1-\alpha$.6 Due to the discrete nature of empirical quantiles, the actual coverage is often slightly conservative (greater than $1-\alpha$).33

Key Variants

Two main practical implementations of CP exist:

Full/Transductive CP (TCP): This is the original formulation.[6] To form the prediction set for a new test point Xtest, TCP considers each possible label $y \in Y$. For each y, it temporarily adds (Xtest,y) to the training data, retrains the underlying model and nonconformity function on this augmented dataset, calculates the nonconformity scores for all points (including the test point), computes the p-value for (Xtest,y), and includes y in the set if $p(y)>\alpha$. This process requires retraining the model $|Y|$ times for *each* test point, making it computationally prohibitive for modern ML models (like deep networks) and large datasets or label spaces.[34]

Split/Inductive CP (ICP): To overcome the computational burden of TCP, the split or inductive approach was developed.[6] The available labeled data is split into two disjoint sets: a **proper training set** (Dtrain) and a **calibration set** (Dcal).

1. The underlying model (e.g., a neural network classifier) is trained only once on Dtrain.
2. The trained model is then used to compute nonconformity scores for all points in the separate calibration set Dcal.
3. The quantile threshold $\hat{q}$ is determined from these calibration scores.

4. For a new test point Xtest, the prediction set is formed by evaluating s(Xtest,y) for all possible y using the single trained model and comparing the scores to q^. ICP is vastly more computationally efficient as the model training happens only once.[27] However, splitting the data means less data is available for training the model and less data for calibration, which can lead to reduced statistical efficiency (potentially wider prediction sets) compared to TCP.[6] **Mondrian Conformal Prediction** is a specific type of ICP that allows for defining different nonconformity measures or calibrating separately for predefined, disjoint subgroups of the data, enabling exact group-conditional coverage for those groups.[7]

3. Addressing Conformal Prediction's Efficiency: The Role of Test-Time Augmentation (TTA)

The Challenge: Impractically Large Prediction Sets

While the marginal coverage guarantee of CP is a powerful theoretical property, a significant practical limitation is that the resulting prediction sets can often be uninformatively large.[24] In complex classification problems, such as medical image analysis with numerous potential diagnoses or general image recognition tasks like ImageNet (with 1000 classes), standard CP methods might output sets containing dozens or even hundreds of labels.[42] For a clinician presented with a list of 200 possible conditions for a patient based on an X-ray, the prediction set offers little practical guidance, despite its statistical validity.[53] This issue arises because the calibration process must be conservative enough to guarantee coverage across all possible inputs, and if the underlying model's predictions are not sharp or well-calibrated, the resulting nonconformity score threshold (q^) becomes large, leading to inclusive sets.

Test-Time Augmentation (TTA): Concept and General Application

Test-Time Augmentation (TTA) is a technique commonly employed in computer vision to enhance the performance of trained models during the inference phase.[40] The core idea is to create multiple slightly modified versions (augmentations) of a single test input image. These augmentations are typically label-preserving transformations like random cropping, horizontal flipping, rotation, zooming, or adjustments to brightness and contrast.[40] The pre-trained model is then run on each of these augmented versions, generating multiple predictions for the original input. These predictions are subsequently aggregated (e.g., by averaging the predicted probabilities) to produce a final, often more robust and accurate, prediction.[40] TTA helps to smooth out model sensitivities to minor input variations and effectively averages out noise, leading to

improved accuracy, robustness against perturbations, and better calibration of predictive scores.[40]

MIT's TTA-Enhanced Conformal Prediction

Recognizing the potential of TTA to improve the quality of predictions fed into the CP framework, researchers at MIT (Shanmugam et al.) developed a method combining TTA with conformal prediction (TTA-CP) to address the large prediction set problem.[40]

Methodology: The TTA-CP approach integrates TTA as a pre-processing step before conformal calibration and prediction.[40]

1. **Data Splitting:** The available labeled data (beyond the initial model training set) is divided into two disjoint sets: a set for learning the TTA aggregation policy (DTTA) and a set for conformal calibration (Dcal).
2. **TTA Application:** For each image x (in DTTA, Dcal, or at test time), multiple augmented versions $a_1(x),\ldots,a_m(x)$ are generated using a predefined augmentation policy $A=\{a_0,\ldots,a_{m-1}\}$ (where a_0 is often the identity).[41]
3. **Prediction Generation:** The underlying pre-trained model f generates predictions (e.g., probability vectors) for each augmented image: $f(a_j(x))$.
4. **Learning Aggregation:** An aggregation function g (e.g., a weighted average of probability vectors) is learned using the data in DTTA. The goal is to find the aggregation strategy that maximizes the accuracy or another desired metric of the aggregated predictions on DTTA.[40]
5. **Conformal Calibration:** The learned TTA policy (augmentations A and aggregation g) is applied to the calibration set Dcal. Nonconformity scores $\alpha_i=s(X_i,Y_i)$ are calculated based on the *aggregated* TTA predictions for each $(X_i,Y_i)\in D_{cal}$. The conformal quantile threshold $\hat{q}$ is then computed from these scores based on the desired error rate α.[41] The specific score used in the MIT work was based on Romano et al.'s method, related to the cumulative probability needed to include the true class.[41]
6. **Prediction Set Construction:** For a new test image X_{test}, TTA is applied, predictions are aggregated using the learned policy g, and the final prediction set $C(X_{test})$ is formed by including all labels y whose nonconformity score $s(X_{test},y)$ (calculated using the aggregated prediction) is below the threshold $\hat{q}$.[41]

Crucially, this entire process requires **no retraining** of the original underlying model f.[40] It acts purely as a post-processing wrapper.

Preserving Validity: A key aspect is maintaining the validity of the conformal guarantee. By using separate, disjoint datasets for learning the TTA aggregation policy

(DTTA) and for conformal calibration (Dcal), the method ensures that the calibration data remains exchangeable with the test data, conditioned on the (fixed) learned TTA policy. The TTA transformation is applied identically to calibration and test points, preserving the conditions needed for the $P(Y \in C(X)) \geq 1-\alpha$ guarantee to hold.[41]

Key Findings

The experimental results of the TTA-CP method demonstrated significant improvements:

- **Set Size Reduction:** Compared to standard conformal prediction methods across several image classification benchmarks (including ImageNet, iNaturalist, CUB-Birds), TTA-CP reduced the average prediction set sizes substantially, with reductions ranging from **10% to 30%.**[40] This makes the output sets considerably more informative and potentially actionable.
- **Coverage Maintenance:** This significant gain in efficiency (smaller sets) was achieved **without compromising the theoretical marginal coverage guarantee**.[40] The prediction sets still contained the true label with the pre-specified probability (e.g., ≥95% if $\alpha=0.05$).
- **Data Allocation Trade-off:** The research revealed an interesting finding regarding data usage. Even though some labeled data was "sacrificed" from the calibration pool to create DTTA for learning the aggregation policy, the resulting improvement in the quality of the aggregated predictions was substantial enough to **outweigh the potential loss in calibration power** from having a slightly smaller Dcal.[40] This suggests that strategically allocating labeled data for post-training refinement steps like TTA can be more beneficial than using all available data solely for calibration, raising important questions for future work on optimal post-training data utilization.[40]
- **Impact on Low-Confidence Classes:** An analysis revealed that TTA helps improve prediction set sizes partly because it increases the predicted probability (or improves the rank) of the true class even when the base model initially predicts it as very unlikely (e.g., ranking it 200th). While this might not change the top prediction, it significantly impacts the calculation of cumulative nonconformity scores used in CP, leading to smaller sets.[41]

The effectiveness of TTA-CP stems from its ability to enhance the underlying predictions before they are conformalized. By making the model's outputs more accurate, robust to minor input variations, and potentially better calibrated through augmentation and aggregation, TTA provides a stronger foundation for the conformal procedure. This allows the calibration step to derive a tighter threshold $\hat{q}$, resulting in smaller prediction sets without violating the coverage guarantee. This demonstrates

that improving the *quality* of the base predictions is a powerful lever for improving the *efficiency* of conformal prediction.

Furthermore, TTA-CP offers a practical advantage. As it requires no modification to the original model training process and utilizes standard data augmentation techniques, it serves as a relatively simple "plug-and-play" enhancement.[40] This contrasts sharply with other approaches that might necessitate complex model retraining (like conformal training [56]) or adopting entirely different architectures (like Bayesian Neural Networks). The accessibility of TTA-CP lowers the barrier for practitioners seeking to obtain more useful and informative uncertainty guarantees from their existing models, potentially accelerating the adoption of reliable AI in fields like medicine.

4. Beyond Marginal Coverage and Set Size: Advanced Conformal Methods

The Need for Adaptivity

While the marginal coverage guarantee $P(Y \in C(X)) \geq 1-\alpha$ is the defining feature of standard CP, its limitation lies in the "marginal" aspect. The guarantee holds on average over the entire data distribution but provides no assurance about performance for specific subgroups or individual instances.[6] A model might achieve 95% average coverage overall, but systematically fail to cover the true diagnosis for a particular patient demographic or for instances with specific challenging features.[48] In high-stakes medical decision-making, such disparities are unacceptable; reliability is needed at a more granular level.[7] However, achieving exact, distribution-free coverage *conditional* on every possible input $X=x$ is theoretically impossible in finite samples without making stronger assumptions.[36] This has motivated the development of advanced CP methods aiming for stronger, more adaptive forms of guarantees.

Conditional Conformal Prediction

The goal of conditional conformal prediction is to achieve coverage guarantees that hold conditionally on specific features, groups, or properties of the data points.[36]

- **Group-Conditional CP (Mondrian CP):** As mentioned earlier, Mondrian CP partitions the data into predefined, *disjoint* groups (e.g., based on patient age categories, hospital site) and performs calibration separately within each group.[7] This yields prediction sets $C_g(X)$ for each group g such that $P(Y \in C_g(X) \mid X \in \text{group } g) \geq 1-\alpha$. While providing exact conditional coverage for these specific groups, it doesn't handle overlapping groups or provide guarantees conditional on continuous features.[37]

- **Approximate Conditional Coverage:** Recognizing the impossibility of exact point-wise conditional coverage, many methods aim for *approximate* conditional coverage. This involves strategies like ensuring coverage holds conditional on certain statistics derived from the input or model output, rather than the full input X. For example, researchers have proposed methods targeting coverage conditional on the model's confidence level and a "trust score" measuring deviation from the ideal Bayes classifier, aiming to ensure coverage even for overconfident incorrect predictions.[48] Label-conditional CP calibrates thresholds separately for each possible output label, which can be useful in epidemiological surveillance from EHR text.[57]

Locally Adaptive Conformal Prediction & CQR

A major focus has been on making prediction intervals (in regression) or set sizes (in classification) adaptive to the local characteristics of the input space, particularly for **heteroscedastic** data where the level of noise or uncertainty varies depending on the input features X.

- **Locally Adaptive CP:** These methods aim to adjust the size of the prediction interval/set based on an estimate of the local variability or difficulty. A common approach involves defining a nonconformity score that normalizes the prediction error by an estimate of the local scale of errors, often the conditional mean absolute deviation (MAD) $\sigma\hat{}(x)=E[|Y-\mu\hat{}(x)||X=x]$.[44] The score becomes $s(x,y)=|y-\mu\hat{}(x)|/\sigma\hat{}(x)$. This makes the threshold $q\hat{}$ effectively scaled by $\sigma\hat{}(x)$ when forming the interval, leading to wider intervals where local variability is high and narrower intervals where it is low.[45] Techniques like gradient boosting have been proposed to systematically learn and optimize these adaptive score functions based on desired properties like minimizing average interval length while maintaining coverage.[44]
- **Conformalized Quantile Regression (CQR):** CQR offers an elegant way to achieve adaptivity by directly leveraging quantile regression.[44] Quantile regression models estimate conditional quantiles (e.g., the 5th and 95th percentiles for a 90% interval) directly, naturally capturing heteroscedasticity.[46] CQR works as follows:
  1. Train quantile regression models on Dtrain to get estimates of the lower ($q\hat{}\alpha/2(x)$) and upper ($q\hat{}1-\alpha/2(x)$) conditional quantiles.
  2. On the calibration set Dcal, compute conformity scores based on the signed distance of the true Yi from the estimated interval: $Ei=\max(q\hat{}\alpha/2(Xi)-Yi, Yi-q\hat{}1-\alpha/2(Xi))$. A score $Ei \le 0$ means Yi is within the interval.

3. Calculate the (1−α) quantile q^ of these calibration scores Ei.
4. Construct the final interval for a new Xtest by adjusting the initial quantile interval: C(Xtest)=[q^α/2(Xtest)−q^,q^1−α/2(Xtest)+q^].[46] CQR inherits the adaptivity of quantile regression and the rigorous finite-sample coverage guarantee of conformal prediction.[46] It has shown strong empirical performance, often producing shorter intervals than other conformal methods, especially for heteroscedastic data.[44]

Conformal Risk Control (CRC)

Standard CP focuses on controlling the miscoverage rate (Type I error). Conformal Risk Control (CRC) generalizes this framework to control other, potentially more relevant, risk metrics.[5] In medical diagnosis, controlling the **False Negative Rate (FNR)** is often critical, as missing a disease can have severe consequences. CRC provides methods to construct prediction sets such that the chosen risk (e.g., FNR) is guaranteed to be below a user-specified level α.

- **Conformal Risk Adaptation (CRA):** Applying standard CRC to tasks like image segmentation can still result in poor *conditional* risk control – the average FNR might be controlled, but some images might suffer very high FNR while others have almost none.[36] CRA was developed to address this for segmentation.[36] It introduces a novel score function based on adaptive prediction sets (similar to those used in classification that capture a certain mass of predicted probability). This allows the prediction threshold to adapt to the confidence distribution of each image, leading to significantly improved conditional risk control (e.g., more consistent FNR across different images) while maintaining the marginal risk guarantee.[36] CRA builds on a theoretical connection established between CRC and CP via a weighted quantile approach.[36]

Specialized Frameworks

- **Conformal Triage:** This algorithm provides a practical framework for deploying AI in clinical workflows, specifically for medical imaging.[38] Instead of outputting sets, it triages patients into three categories:
  - **Low-Risk:** Guaranteed high Negative Predictive Value (NPV ≥1−αNPV).
  - **High-Risk:** Guaranteed high Positive Predictive Value (PPV ≥1−αPPV).
  - **Uncertain:** Cases where the desired PPV/NPV guarantee cannot be met, requiring human review. The thresholds for categorization are determined using conformal calibration on a local dataset, making the guarantees robust to distribution shifts between the model's training data and the deployment site.[38]

- **Utility-Directed CP:** This line of research aims to make prediction sets more directly useful for downstream tasks by incorporating a notion of decision utility into their construction.[28] Instead of optimizing solely for statistical coverage or minimal size, the goal is to generate sets that are "actionable." For example, in diagnosis, a set might be considered more useful if all diagnoses within it share the same optimal treatment plan, or if they can be easily distinguished with further low-cost tests.[28] This moves beyond purely statistical guarantees towards aligning uncertainty quantification with practical decision-making objectives.

The development and diversification of these advanced CP methods indicate a significant maturation of the field. Initial work established the foundational guarantees of marginal coverage.[4] However, practical deployment, especially in sensitive domains like medicine, quickly revealed the limitations of average guarantees.[7] The subsequent focus on conditional coverage [36], local adaptivity [44], control of specific clinical risks [5], and alignment with decision utility [28] demonstrates a clear trajectory towards tailoring CP's rigorous guarantees to meet complex, real-world requirements. CP is evolving from a purely statistical tool into a flexible framework for generating nuanced and practically relevant reliability assurances.

However, this advancement comes with inherent trade-offs. While standard marginal CP relies only on the relatively weak exchangeability assumption [4], achieving stronger, more localized guarantees often requires additional steps or assumptions. Exact conditional coverage is impossible without distributional assumptions.[36] Group-conditional CP necessitates defining meaningful, often disjoint, groups.[48] Locally adaptive methods like CQR or those using MAD estimates involve fitting secondary models (for quantiles or variance), introducing their own modeling choices and potential sources of error.[44] Methods aiming for approximate conditional coverage might rely on specific heuristics or model properties that may not always hold.[48] Practitioners must therefore carefully balance the desire for stronger, more granular guarantees against the increased complexity, data requirements, and potential fragility of the methods needed to achieve them.

5. Conformal Prediction in the Landscape of Uncertainty Quantification

Comparison Overview

Conformal prediction is one of several approaches available for quantifying uncertainty in machine learning models. Understanding its unique characteristics requires comparing it to other prevalent methods, particularly Bayesian inference, MC

Dropout, and Deep Ensembles, which are frequently used in medical AI.[11] Each method offers different types of uncertainty information, relies on different assumptions, provides different guarantees (if any), and involves different computational trade-offs.

Conformal Prediction (CP)

- **Core Approach:** A post-hoc "wrapper" method that uses a calibration dataset to convert point predictions from any underlying model into prediction sets (classification) or intervals (regression). Relies on nonconformity scores to measure the "strangeness" of potential predictions relative to calibration data.[11]
- **Key Assumptions:** Data exchangeability (between calibration and test sets).[4]
- **Type of Guarantee:** Finite-sample, distribution-free marginal coverage guarantee ($P(Y_{true} \in C(X)) \geq 1-\alpha$).[4]
- **Output:** Prediction set or interval.[11]
- **Pros:** Rigorous theoretical guarantee, distribution-free, model-agnostic, computationally efficient (for ICP).[4]
- **Cons:** Guarantee is marginal (not conditional by default), efficiency (set size) depends heavily on the choice of NCS and underlying model quality, requires a separate calibration set (for ICP).[24]
- **Medical Applicability:** Widely applicable across diagnosis, prognosis, segmentation, genomics, drug discovery where provable reliability is desired.[5]

Bayesian Inference (e.g., Bayesian Neural Networks)

- **Core Approach:** Treats model parameters (e.g., network weights) as random variables with prior distributions. Uses Bayes' theorem to compute posterior distributions of parameters given observed data. Predictions are made by averaging over the posterior distribution, yielding predictive distributions that capture uncertainty.[4]
- **Key Assumptions:** Correctness of the model structure and the specified prior distributions.[4]
- **Type of Guarantee:** Guarantees are typically asymptotic or rely on the model assumptions being correct. No finite-sample, distribution-free coverage guarantee like CP.[4]
- **Output:** Full posterior predictive distribution (or samples from it).[14]
- **Pros:** Principled framework for incorporating prior knowledge, can distinguish between epistemic and aleatoric uncertainty, provides rich distributional information.[11]
- **Cons:** Computationally intensive (e.g., MCMC sampling), requires specifying priors, validity hinges on assumptions, approximations (like Variational Inference)

may be needed and introduce their own errors.[4]
- **Medical Applicability:** Used for uncertainty estimation in various tasks, especially where incorporating prior clinical knowledge is beneficial or detailed uncertainty decomposition is needed.[9]

Monte Carlo (MC) Dropout

- **Core Approach:** An approximation technique for Bayesian inference in neural networks. Applies dropout layers not only during training but also during multiple forward passes at test time. The variability (e.g., variance) of the predictions across these passes is used as an estimate of uncertainty.[11]
- **Key Assumptions:** Assumes dropout can approximate Bayesian model averaging; relies on the network architecture including dropout layers.
- **Type of Guarantee:** No rigorous theoretical guarantees on coverage or calibration, provides a heuristic measure of uncertainty.[23]
- **Output:** Typically mean prediction and variance (or samples).[14]
- **Pros:** Relatively simple to implement on existing networks with dropout, computationally cheaper than full Bayesian methods or ensembles.[14]
- **Cons:** Provides only an approximation of uncertainty, quality of estimate depends on dropout rate and network specifics, lacks theoretical guarantees.[12]
- **Medical Applicability:** Popular practical method for UQ in deep learning for medical imaging (segmentation, classification) due to ease of implementation.[14]

Deep Ensembles

- **Core Approach:** Trains multiple (typically 5-10) identical network architectures independently from different random initializations on the same training data. At test time, predictions from all models are aggregated (e.g., averaged). The disagreement (e.g., variance) among the ensemble members' predictions is used as a measure of uncertainty.[9]
- **Key Assumptions:** Assumes diversity among ensemble members captures model uncertainty.
- **Type of Guarantee:** No rigorous theoretical guarantees on coverage, provides an empirically strong heuristic for uncertainty.[12]
- **Output:** Mean prediction and variance (or samples).[12]
- **Pros:** Empirically shown to produce high-quality predictions and reliable uncertainty estimates, often outperforming MC Dropout and approximate Bayesian methods.[12] Conceptually simple.
- **Cons:** Computationally very expensive due to training and storing multiple independent models.[12]
- **Medical Applicability:** Used when high empirical performance is critical and

computational resources allow, e.g., in critical diagnostic tasks.[14]

Comparative Analysis Table

Feature	Conformal Prediction (CP)	Bayesian Inference	Monte Carlo (MC) Dropout	Deep Ensembles
Core Approach	Post-hoc calibration via nonconformity scores	Compute posterior over parameters	Approximate Bayesian via dropout	Train & average multiple models
Key Assumptions	Data exchangeability	Correct model & priors	Dropout approximates Bayesian avg.	Ensemble diversity captures uncertainty
Type of Guarantee	Finite-sample, distribution-free marginal coverage	Asymptotic or assumption-dependent	None (heuristic)	None (heuristic)
Output	Prediction Set / Interval	Posterior predictive distribution	Mean & Variance / Samples	Mean & Variance / Samples
Pros	Rigorous guarantee, model-agnostic, flexible	Principled, prior knowledge, uncertainty decomposition	Simple implementation, relatively fast	Strong empirical performance
Cons	Marginal guarantee, efficiency depends on NCS	Computationally expensive, prior sensitivity	Approximation quality varies, no guarantee	Very computationally expensive
Medical Applicability	Diagnosis, segmentation, risk prediction where guarantees needed	Tasks needing prior knowledge, detailed UQ	Common practical UQ for DL models	High-stakes tasks if compute allows

Synergies and Distinctions

CP stands apart due to its distribution-free guarantee, which does not depend on the correctness of the underlying model or strong distributional assumptions, unlike Bayesian methods.[4] This robustness is a major appeal in medicine, where data complexities often violate the assumptions required by probabilistic methods. While distinct, CP is not entirely isolated. The nonconformity score in CP can be derived from the outputs of any model, including Bayesian models or ensembles. For instance, the uncertainty estimate from a Bayesian model could potentially be used to define a more effective, adaptive nonconformity score for CP [27], although the final guarantee would still stem from the conformal procedure itself.

The fundamental trade-off often lies between the nature of the guarantee and the richness of the information provided. CP offers a specific, robust guarantee: the true outcome lies within the prediction set with at least probability $1-\alpha$.[11] This guarantee, however, provides limited information about the *distribution* of likelihood within the set. Bayesian methods, conversely, aim to provide a full posterior distribution, offering potentially richer insights into the uncertainty structure (e.g., multimodality).[22] However, the validity of this richer information is contingent on the strong assumptions underlying the Bayesian model.[4] The choice between these paradigms depends on whether the application prioritizes a rigorous, albeit potentially less detailed, coverage guarantee (CP) or richer, more detailed uncertainty information that comes with stronger, potentially unverifiable assumptions (Bayesian methods). MC Dropout and Ensembles offer practical heuristics that often work well empirically but lack the formal guarantees of either CP or (under correct assumptions) Bayesian inference.

6. Applications of Conformal Prediction in Medical Diagnosis

Overview

The model-agnostic nature and rigorous guarantees of conformal prediction have spurred its application across a diverse range of medical domains and data modalities. Its ability to provide reliable uncertainty estimates is proving valuable for tasks ranging from image analysis and genomic interpretation to clinical risk prediction and drug development.[4]

Medical Imaging Analysis

Medical imaging is a prime area for CP application, given the visual nature of

diagnosis and the inherent ambiguity in many images.

- **Radiology/Pathology Diagnosis:** CP can enhance diagnostic workflows by providing clinicians with a set of differential diagnoses instead of a single AI prediction, particularly for ambiguous cases.[17] For instance, distinguishing between pleural effusion and pulmonary infiltrates on a chest X-ray, which can appear similar, is a scenario where a prediction set {pleural effusion, pulmonary infiltrate} with a high confidence guarantee could be more useful than a single, potentially incorrect, label.[40] The Conformal Triage algorithm specifically uses CP principles to categorize head CT scans into high-risk (high PPV guaranteed), low-risk (high NPV guaranteed), and uncertain groups, facilitating workflow management and ensuring reliability even under distribution shift.[38] Utility-directed CP aims to make these sets even more actionable, potentially grouping diagnoses by treatment implications.[28]
- **Segmentation Uncertainty:** Accurately segmenting regions of interest (e.g., tumors, organs) is crucial for treatment planning and monitoring. CP, CRC, and CRA methods are being used to quantify pixel-level uncertainty in segmentations.[13] A notable example involves using Mondrian ICP to assess uncertainty in deep learning-based prostate segmentation on MRI.[13] By identifying and excluding pixels with uncertain classifications (based on an 85% confidence level), the researchers significantly improved the accuracy and reliability of prostate volume measurements, showing better agreement with the reference standard compared to the raw DL output.[13] Similarly, CRA has been applied to polyp segmentation to provide more consistent control over false negative rates across different images.[37]

Genomic Medicine

The complexity and high dimensionality of genomic data make uncertainty quantification essential. CP offers a promising approach.[6]

- **Variant Calling/Prioritization:** CP can provide confidence sets for genetic variants identified through sequencing, aiding in distinguishing true mutations from noise and prioritizing variants for further investigation in disease diagnosis.
- **Pharmacogenomics:** Predicting individual drug responses based on genomic profiles is a key goal of personalized medicine. CP can generate prediction intervals for drug sensitivity or classify patients as likely responders/non-responders with guaranteed error rates, improving the safety and reliability of treatment selection.[6]
- **Immunotherapy Response:** Predicting response to treatments like immune checkpoint inhibitors often relies on biomarkers like tumor mutational burden. CP

can quantify the uncertainty associated with these predictions.
- **Antimicrobial Resistance:** CP can provide reliable predictions of whether a pathogen is resistant to specific antibiotics based on its genomic features, supporting crucial treatment decisions.

Clinical Risk Prediction

Predicting patient risk based on clinical data is another critical application area.

- **Sepsis Mortality and Diagnosis:** Studies have successfully applied CP to predict in-hospital mortality risk for sepsis patients in the ICU.[7] Using Mondrian CP with a gradient boosting model trained on EHR data, the system (CPMORS) provided risk predictions along with confidence levels, flagging uncertain cases for clinician review. This approach significantly reduced the error rate compared to the base model and outperformed traditional scoring systems.[7] Another study used CP with deep learning on time-series data for *early* sepsis prediction in non-ICU patients, demonstrating high accuracy and improved specificity by reducing false positives via the conformal framework.[61]
- **Bacterial Infection Focus:** CP combined with ML models using routine biochemical data and vital parameters from EHRs has been used to predict the site of bacterial infection (e.g., airway, urine, blood) with calibrated confidence estimates, potentially speeding up diagnosis.[5]
- **Disease Outbreak Surveillance:** Preliminary work explores using label-conditional CP with active learning on unstructured EHR text to automate the detection of emerging disease patterns.[57]

Drug Discovery and Development

CP is also utilized in the pharmaceutical pipeline.[12] Applications include predicting molecular properties, screening potential drug candidates, assessing toxicology risks, predicting pharmacokinetic profiles, and identifying potential drug targets, all with associated confidence levels provided by the CP framework.[26]

Other Applications

The versatility of CP is further demonstrated by its application in diverse areas such as syndrome differentiation in Traditional Chinese Medicine using multi-label CP with Random Forests [25], estimating depression severity from facial videos with confidence intervals [50], and providing reliable answer sets for medical multiple-choice question-answering by large language models.[35]

The wide array of applications underscores CP's inherent flexibility. Its model-agnostic

nature allows it to be readily applied to various algorithms, from traditional ML models like Random Forests [25] and Gradient Boosting [5] to complex deep learning architectures [13] and even large language models.[35] Furthermore, its applicability spans diverse data types encountered in medicine, including images [13], genomic sequences [6], structured EHR and biochemical data [5], time-series data [61], textual data [35], and chemical structure data for drug discovery.[12] This adaptability makes CP a strong candidate for a universal UQ framework wherever statistically rigorous reliability guarantees are paramount in healthcare settings.

However, a closer look at these successful applications reveals that achieving practical utility often requires more than just applying basic CP. Many studies employ sophisticated base models (DL, XGBoost) known for high predictive performance.[5] Furthermore, they often utilize advanced CP variants tailored to the specific problem: Mondrian CP for handling heterogeneity in classification [7], CRC or CRA for controlling specific risks like FNR in segmentation [36], CQR for adaptive regression intervals [46], or specialized frameworks like Conformal Triage.[38] The need for enhancements like TTA to improve efficiency (Section 3) further supports this observation. This suggests that while CP provides a robust theoretical foundation, realizing its full potential in complex medical domains often necessitates a synergistic approach: combining it with powerful underlying predictive models and carefully selecting or developing advanced CP methodologies (e.g., adaptive nonconformity scores, conditional guarantees, risk control) that address the specific nuances and requirements of the clinical task. Basic CP might serve as a starting point, but tailored adaptations are frequently key to practical success.

7. Bridging the Gap: Clinical Integration and Decision Support

Interpreting Conformal Prediction Sets

While CP provides statistically sound prediction sets, translating these sets into actionable clinical insights presents a significant challenge.[16] A prediction set like {Eczema, Psoriasis} [28] or {Pneumonia, Pulmonary Embolism, Congestive Heart Failure} raises questions for the clinician: How should this information be used? Does it effectively narrow the diagnostic possibilities, or does it increase cognitive burden by presenting multiple options?.[32]

Several interpretation strategies exist. Clinicians might use the set to confidently rule out diagnoses not included within it. They could focus their differential diagnosis process on the conditions listed in the set. The size of the prediction set itself can serve as an intuitive measure of uncertainty – a smaller set implies higher confidence.[62] However, there is often a gap between the statistical guarantee of

coverage (the true diagnosis is likely in the set) and the set's direct utility for making a specific decision.[16] For instance, a set might be statistically valid but clinically unhelpful if the included diagnoses require vastly different or conflicting treatments. Utility-directed CP aims to bridge this gap by constructing sets aligned with downstream decision preferences.[28]

Furthermore, human decision-making is complex. Clinicians bring their own prior knowledge, experience, and potentially private information about the patient that is not available to the AI model.[32] How they integrate a prediction set with this existing knowledge is not straightforward. Research involving human participants suggests that interaction with prediction sets can have nuanced and sometimes counterintuitive effects, even impacting the fairness of decisions.[51] Simply providing a statistically valid set does not guarantee optimal or even improved decision-making.

Potential Pathways for Clinical Workflow Integration

Despite interpretation challenges, several pathways exist for integrating CP into clinical workflows:

- **Decision Support and Case Flagging:** CP can act as a safety net by identifying cases where the AI model is uncertain. Prediction sets containing multiple disparate diagnoses, or simply large sets, can automatically flag cases for mandatory review by a human expert.[7] The CPMORS sepsis mortality predictor, for example, used the conformal output to identify uncertain predictions requiring clinician attention; these uncertain cases were later found to correlate with higher rates of complications like acute kidney injury.[7]
- **Triage Systems:** Algorithms like Conformal Triage offer a structured way to integrate CP into workflows by stratifying patients based on guaranteed risk levels (high PPV for high-risk, high NPV for low-risk), directing clinician attention to the high-risk and uncertain groups.[38]
- **Safety Layer for AI Deployment:** CP can serve as a crucial validation and safety layer applied to existing AI tools before clinical deployment. By wrapping a pre-trained model with CP, institutions can gain assurance about its reliability under specific error tolerance levels.[6]
- **Integration with EHRs and Imaging Systems:** Seamless integration into existing hospital IT infrastructure, such as EHRs and Picture Archiving and Communication Systems (PACS), is essential for practical adoption, allowing CP outputs to be presented directly within the clinician's standard workflow.[1]

Challenges to Clinical Adoption

Several hurdles remain for the widespread clinical adoption of CP:

- **Data Requirements:** ICP requires a dedicated calibration set. For guarantees to be meaningful in a specific clinical environment, this calibration data should ideally be representative of the local patient population and data acquisition practices, potentially requiring ongoing local data collection and calibration.[6] The data splitting in ICP also reduces the amount of data available for model training, which can be a limitation when labeled medical data is scarce or expensive (data inefficiency).[6]
- **Computational Cost:** While ICP is far more efficient than TCP, advanced methods can add overhead. TTA requires multiple model inferences per test case.[40] Ensembles used as base models are inherently costly.[15] Real-time application in busy clinical settings demands computationally lean solutions.
- **User Trust and Training:** Clinicians need to understand what conformal prediction sets represent, what the 1–α guarantee means (and what it doesn't mean, e.g., conditional coverage), and how to incorporate this information into their decision-making.[7] Effective training and clear, intuitive presentation of CP outputs are crucial to build trust and avoid misuse.[1] Overcoming automation bias – the tendency to over-rely on automated systems – is also critical, and CP's ability to flag uncertainty can help mitigate this.[7]
- **Choosing Alpha (α):** Selecting the appropriate significance level α (e.g., 0.05 for 95% confidence) is a critical decision. A smaller α provides a stronger guarantee but typically leads to larger, potentially less useful, prediction sets. The optimal choice depends on the clinical context, the tolerance for error, and requires domain expertise.[6]

The successful integration of CP into clinical practice hinges on more than just its statistical properties. While validity is foundational, usability is paramount. This requires a human-centered design approach, considering how clinicians perceive, interpret, and act upon prediction sets within the constraints of their demanding workflows.[1] Research on human-AI interaction, development of intuitive visualization tools for prediction sets, and robust user training programs are as important as algorithmic advancements.[1] Without careful attention to these human factors, even statistically impeccable CP methods may fail to deliver their intended benefits in real-world clinical settings.

Furthermore, the requirement for local calibration data presents both a challenge and an opportunity. While managing local calibration datasets adds an operational burden for healthcare institutions [38], it directly addresses a major weakness of many deployed AI systems: poor generalization due to distribution shifts between the original training

data and the local deployment environment.[7] By calibrating locally, CP can provide reliability guarantees that are specifically tailored and validated for the context in which the AI is being used. This localization can significantly enhance trustworthiness and provide a pathway for deploying AI systems that are demonstrably reliable *at the point of care*.

8. Ethical and Regulatory Dimensions

The deployment of AI, including methods like conformal prediction, in medical diagnosis carries significant ethical and regulatory implications that must be carefully navigated.

Fairness Considerations

Ensuring fairness across different patient populations is a critical ethical requirement. While CP provides marginal coverage guarantees, these average guarantees do not preclude disparities in performance across demographic groups (e.g., based on race, sex, age).[36]

- **Disparate Impact:** A key concern is that prediction sets might be systematically larger or less accurate for certain protected groups, even if the overall 1–α coverage is met. If an underlying model is less accurate for a specific group, standard CP might produce larger, less informative sets for individuals in that group to maintain the coverage guarantee. This difference in utility can lead to disparate impact in downstream decisions.
- **Equalized Coverage vs. Fairness Outcomes:** A common approach to algorithmic fairness is to strive for **Equalized Coverage**, meaning the 1–α guarantee holds conditionally for each predefined group. Methods like Mondrian CP can achieve this for disjoint groups.[48] However, experiments involving human participants making decisions based on conformal sets revealed a troubling finding: providing sets satisfying Equalized Coverage (which often involves varying set sizes across groups) actually *increased* disparate impact in the humans' decisions compared to using sets derived from standard marginal CP.[51] This suggests that humans may react differently to prediction sets of varying sizes, potentially trusting larger sets less or finding them harder to use, leading to biased outcomes even when the statistical fairness metric is satisfied. The study proposed that aiming for **Equalized Set Size** might be a better heuristic for reducing downstream disparate impact, but this highlights a critical disconnect. Simply enforcing statistical fairness criteria on the AI output does not guarantee fairness in the overall human-AI decision-making process. A more holistic, end-to-end evaluation considering human interaction is necessary to develop

truly fair systems.

Safety, Trustworthiness, and Automation Bias

CP directly contributes to enhancing the safety and trustworthiness of medical AI.[6] By providing explicit bounds on uncertainty and guaranteeing coverage, it offers a more reliable alternative to opaque black-box predictions or heuristic confidence scores. This explicit uncertainty quantification can help mitigate **automation bias**, the tendency for humans to excessively rely on automated recommendations.[7] When CP outputs a large prediction set or flags a case as uncertain (e.g., in Conformal Triage), it serves as a clear signal to the clinician that human expertise and careful review are required, rather than blindly accepting an AI's single best guess.[7] Transparency regarding how uncertainty estimates are generated and what the guarantees mean is crucial for building and maintaining clinician trust.[1]

Regulatory Perspectives (FDA)

Regulatory bodies like the U.S. Food and Drug Administration (FDA) are actively developing frameworks to oversee the safe and effective use of AI/ML in healthcare.

- **AI/ML as SaMD:** The FDA regulates AI/ML software intended for medical purposes as Software as a Medical Device (SaMD), applying a risk-based classification system (Categories I-IV) based on intended use and potential impact on patient health.[17] Higher-risk applications like diagnosis typically require more stringent review pathways (e.g., 510(k) clearance or Premarket Approval).[17]
- **Challenges for Adaptive AI:** A key challenge recognized by the FDA is the adaptive nature of some AI/ML algorithms, which can learn and change their performance over time based on new data.[3] Traditional device regulation, designed for static devices, struggles with this dynamic behavior.[19]
- **Focus on UQ and Performance Assessment:** The FDA explicitly acknowledges the importance of robust evaluation methods for AI performance and the need for uncertainty quantification.[3] Regulatory science research programs are underway to develop appropriate metrics and tools for assessing AI performance and quantifying uncertainty, with the goal of enabling clinicians to make more informed decisions based on device outputs.[20]
- **Predetermined Change Control Plans (PCCP):** To manage adaptive AI/ML SaMD, the FDA has finalized guidance on Predetermined Change Control Plans.[21] This framework allows manufacturers to pre-specify anticipated modifications to their algorithms (e.g., retraining on new data, adapting parameters) within defined boundaries in their initial regulatory submission. If the changes fall within the approved plan, manufacturers can implement them without requiring a new

submission for each modification, provided they adhere to transparency and real-world performance monitoring commitments.[17] This aims to balance the benefits of iterative improvement with the need for continuous safety and effectiveness oversight.

- **Good Machine Learning Practice (GMLP):** The FDA emphasizes the importance of GMLP, outlining expectations for robust data management, model training methodologies, algorithm interpretability, and rigorous evaluation practices.[17]

The FDA's evolving regulatory landscape, particularly the introduction of PCCPs and the explicit focus on UQ, signals a move towards embracing the potential of dynamic AI systems like those enhanced by CP. However, it also places significant responsibilities on manufacturers. Deploying CP-based systems, especially adaptive ones, will likely require robust protocols for initial calibration, ongoing monitoring of coverage and performance in real-world use, defining appropriate nonconformity scores, managing calibration data, and potentially incorporating CP updates within an approved PCCP. This regulatory evolution creates both opportunities for deploying more reliable AI and challenges related to demonstrating compliance and managing the lifecycle of these sophisticated systems.

9. Conclusion and Future Directions

Recap of Conformal Prediction's Potential

Conformal prediction has emerged as a uniquely valuable framework for uncertainty quantification in the high-stakes domain of medical AI. Its core strengths – model-agnosticism, distribution-free validity, and the provision of rigorous, finite-sample coverage guarantees – directly address the critical need for trustworthy and reliable AI systems in healthcare.[4] By replacing opaque single-point predictions with calibrated prediction sets or intervals, CP offers a principled way to quantify uncertainty and enhance safety in clinical decision support.

Summary of Key Advancements and Applications

Significant progress has been made in enhancing CP's practicality and relevance. Techniques like Test-Time Augmentation (TTA) have proven effective in reducing prediction set sizes, making CP outputs more informative without sacrificing validity.[40] The field has moved beyond basic marginal coverage, developing advanced methods that target approximate conditional coverage, adapt to local data characteristics (e.g., CQR, locally adaptive CP), control specific clinical risks (e.g., CRC, CRA), and even align with downstream decision utility.[28] These advancements have enabled successful applications across diverse medical areas, including radiology, pathology,

genomics, clinical risk prediction, and drug discovery, demonstrating CP's versatility across different data modalities and tasks.[5]

Persistent Challenges

Despite its promise and progress, several challenges remain for the widespread and effective adoption of CP in medicine:

- **Conditional Validity:** Achieving guarantees that hold reliably for specific individuals or subgroups, rather than just on average, remains a major hurdle due to theoretical limitations and practical difficulties.[36]
- **Efficiency:** While methods like TTA help, ensuring prediction sets are consistently small enough to be clinically actionable, especially for complex problems, requires ongoing research into better nonconformity scores and algorithms.[34]
- **Interpretability and Actionability:** Translating statistical prediction sets into clear, intuitive, and useful information for time-constrained clinicians remains a significant barrier.[16]
- **Fairness:** Ensuring that CP methods do not introduce or exacerbate biases across different patient groups, particularly when integrated into human decision-making workflows, requires careful consideration beyond standard statistical metrics.[51]
- **Clinical Integration:** Seamlessly embedding CP into existing clinical workflows and EHR/PACS systems, along with managing local calibration data requirements, poses technical and operational challenges.[1]
- **Regulation:** Navigating the evolving regulatory landscape for adaptive AI/ML SaMD, including demonstrating compliance with frameworks like PCCP, requires significant effort.[17]

Future Research Directions

Addressing these challenges points towards several key areas for future research:

- **Conditional Coverage:** Developing theoretically sound and practically robust methods for achieving meaningful approximate conditional coverage guarantees, perhaps by conditioning on relevant learned features or risk scores.[36]
- **Nonconformity Score Design:** Creating novel nonconformity scores specifically tailored for medical data types (e.g., time-series EHR data, multi-modal data) and clinical objectives, balancing efficiency and validity.[27]
- **Human-AI Interaction:** Designing and evaluating effective methods for presenting conformal prediction outputs to clinicians, studying how they interpret and use this information, and developing interfaces that enhance decision-making.[16]

- **Fairness in Practice:** Moving beyond purely statistical fairness metrics to investigate the end-to-end fairness implications of CP in human-AI collaborative settings and developing methods that promote equitable outcomes.[51]
- **Multimodal and Longitudinal CP:** Extending CP frameworks to effectively handle complex medical data involving multiple modalities (e.g., imaging + genomics + clinical notes) and longitudinal patient trajectories.
- **Validation and Monitoring Standards:** Establishing standardized protocols and best practices for validating CP-based systems in diverse clinical settings and for continuous monitoring of their performance and coverage guarantees post-deployment, aligning with regulatory expectations.[15]
- **Computational Efficiency:** Developing techniques to reduce the computational footprint of advanced CP methods, particularly those involving TTA, ensembles, or complex calibration procedures, to facilitate real-time application.[34]

Concluding Thought

Conformal prediction represents a significant step forward in the quest for reliable and trustworthy AI in healthcare. It is not a panacea, but rather a vital tool in a larger toolkit aimed at ensuring AI systems can be deployed safely and effectively. Its rigorous mathematical foundation provides a level of assurance often missing in other approaches. As research continues to address its practical limitations and explore its potential through advanced methods and careful integration strategies, conformal prediction is poised to play an increasingly crucial role in building a future where AI assists medical professionals with quantified confidence, ultimately benefiting patient care.

Works cited

1. A Survey of Embodied AI in Healthcare: Techniques, Applications, and Opportunities - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2501.07468v1
2. A Review on Revolutionizing Healthcare Technologies with AI and ML Applications in Pharmaceutical Sciences - MDPI, accessed on May 5, 2025, https://www.mdpi.com/2813-2998/4/1/9
3. Artificial Intelligence Program: Research on AI/ML-Based Medical Devices - FDA, accessed on May 5, 2025, https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-program-research-aiml-based-medical-devices
4. arxiv.org, accessed on May 5, 2025, https://arxiv.org/pdf/2410.06494
5. Conformal Prediction and Venn-ABERS Calibration for Reliable Machine Learning-Based Prediction of Bacterial Infection Focus | medRxiv, accessed on

May 5, 2025, https://www.medrxiv.org/content/10.1101/2025.01.21.25320878v2.full
6. Reliable machine learning models in genomic medicine using conformal prediction - PMC, accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11891349/
7. Development and Validation of an Interpretable Conformal Predictor ..., accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10985608/
8. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer, accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11299037/
9. Reliable machine learning models in genomic medicine using conformal prediction, accessed on May 5, 2025, https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2025.1507448/full
10. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches | The BMJ, accessed on May 5, 2025, https://www.bmj.com/content/388/bmj-2024-080749
11. A Comprehensive Guide to Conformal Prediction: Simplifying the Math, and Code, accessed on May 5, 2025, https://daniel-bethell.co.uk/posts/conformal-prediction-guide/
12. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction | Journal of Chemical Information and Modeling - ACS Publications, accessed on May 5, 2025, https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b00975
13. Impact of uncertainty quantification through conformal prediction on ..., accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11607187/
14. A review of uncertainty quantification in medical image analysis ..., accessed on May 5, 2025, https://www.researchgate.net/publication/381319369_A_review_of_uncertainty_quantification_in_medical_image_analysis_Probabilistic_and_non-probabilistic_methods
15. Artificial Intelligence Uncertainty Quantification in Radiotherapy Applications - A Scoping Review - PMC - National Institutes of Health (NIH), accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11118597/
16. Conformal Prediction and Human Decision Making1footnote 11footnote 1We thank John Cherian, Tiffany Ding, and Jason Hartline for feedback on a draft. - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2503.11709v1
17. FDA Review of Radiologic AI Algorithms: Process and Challenges - RSNA Journals, accessed on May 5, 2025, https://pubs.rsna.org/doi/full/10.1148/radiol.230242
18. FDA Review of Radiologic AI Algorithms: Process and Challenges, accessed on May 5, 2025, https://pubs.rsna.org/doi/pdf/10.1148/radiol.230242
19. Managing AI/ML Enabled Medical Devices - Enerxen, accessed on May 5, 2025, https://enerxen.com/2024/01/16/ai-ml-enabled-medical-devices/
20. Evaluation Methods for Artificial Intelligence (AI)-Enabled Medical ..., accessed on May 5, 2025,

https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/evaluation-methods-artificial-intelligence-ai-enabled-medical-devices-performance-assessment-and
21. Regulating radiology AI medical devices that evolve in their lifecycle - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2412.20498v1
22. The challenge of uncertainty quantification of large language models in medicine - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2504.05278v1
23. Conformal Prediction on Quantifying Uncertainty of Dynamic Systems - ResearchGate, accessed on May 5, 2025, https://www.researchgate.net/publication/387104542_Conformal_Prediction_on_Quantifying_Uncertainty_of_Dynamic_Systems
24. Conformalized Link Prediction on Graph Neural Networks - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2406.18763v1
25. Reliable Multi-Label Learning via Conformal Predictor and Random Forest for Syndrome Differentiation of Chronic Fatigue in Traditional Chinese Medicine, accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4053362/
26. CPSign: conformal prediction for cheminformatics modeling - PMC - PubMed Central, accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11214261/
27. A Review of Nonconformity Measures for Conformal Prediction in Regression - Proceedings of Machine Learning Research, accessed on May 5, 2025, https://proceedings.mlr.press/v204/kato23a/kato23a.pdf
28. Utility-Directed Conformal Prediction: A Decision-Aware Framework for Actionable Uncertainty Quantification - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2410.01767v2
29. Conformal Prediction: A Data Perspective - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2410.06494v2
30. jmlr.csail.mit.edu, accessed on May 5, 2025, https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf
31. Confidence on the focal: conformal prediction with selection-conditional coverage - Oxford Academic, accessed on May 5, 2025, https://academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkaf016/8113856?searchresult=1
32. New paper on conformal prediction and human decisions, accessed on May 5, 2025, https://statmodeling.stat.columbia.edu/2025/03/19/new-paper-on-conformal-prediction-and-human-decisions/
33. www.stat.berkeley.edu, accessed on May 5, 2025, https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/conformal.pdf
34. Fast full conformal prediction for multiple test points - AIMS Press, accessed on May 5, 2025, http://aimspress.com/article/doi/10.3934/math.2025236?viewType=HTML
35. Statistical Guarantees of Correctness Coverage for Medical Multiple-Choice Question Answering - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2503.05505

36. Conditional Conformal Risk Adaptation - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2504.07611v1
37. conditional conformal risk adaptation - arXiv, accessed on May 5, 2025, http://www.arxiv.org/pdf/2504.07611
38. (PDF) Conformal Triage for Medical Imaging AI Deployment, accessed on May 5, 2025, https://www.researchgate.net/publication/378166244_Conformal_Triage_for_Medical_Imaging_AI_Deployment
39. (PDF) Assurance monitoring of learning-enabled cyber-physical systems using inductive conformal prediction based on distance learning - ResearchGate, accessed on May 5, 2025, https://www.researchgate.net/publication/351997666_Assurance_monitoring_of_learning-enabled_cyber-physical_systems_using_inductive_conformal_prediction_based_on_distance_learning
40. Making AI models more trustworthy for high-stakes settings | MIT News, accessed on May 5, 2025, https://news.mit.edu/2025/making-ai-models-more-trustworthy-high-stakes-settings-0501
41. dmshanmugam.github.io, accessed on May 5, 2025, https://dmshanmugam.github.io/pdfs/CVPR_2025_TTA_CP.pdf
42. Data Augmentation and Conformal Prediction Helen Lu - DSpace@MIT, accessed on May 5, 2025, https://dspace.mit.edu/bitstream/handle/1721.1/151275/lu-helenl-meng-eecs-2023-thesis.pdf?sequence=1&isAllowed=y
43. CD-split and HPD-split: - Journal of Machine Learning Research, accessed on May 5, 2025, https://www.jmlr.org/papers/volume23/20-797/20-797.pdf
44. NeurIPS Poster Boosted Conformal Prediction Intervals, accessed on May 5, 2025, https://neurips.cc/virtual/2024/poster/95004
45. Boosted Conformal Prediction Intervals - OpenReview, accessed on May 5, 2025, https://openreview.net/pdf?id=Tw032H2onS
46. papers.neurips.cc, accessed on May 5, 2025, http://papers.neurips.cc/paper/8613-conformalized-quantile-regression.pdf
47. Selection by Prediction with Conformal p-values - Journal of Machine Learning Research, accessed on May 5, 2025, https://www.jmlr.org/papers/volume24/22-1176/22-1176.pdf
48. arXiv:2501.10139v2 [cs.LG] 9 Feb 2025, accessed on May 5, 2025, https://arxiv.org/pdf/2501.10139
49. Conformal Prediction Sets with Improved Conditional Coverage using Trust Scores - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2501.10139v2
50. Conformal Depression Prediction - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2405.18723v3
51. Conformal Prediction Sets Can Cause Disparate Impact | OpenReview, accessed on May 5, 2025, https://openreview.net/forum?id=fZK6AQXlUU
52. Enhancing Trustworthiness of AI in Medical Imaging - The Munich Eye, accessed on May 5, 2025,

https://themunicheye.com/boosting-trust-ai-models-medical-imaging-20147
53. MIT Combines Test-Time Augmentation and Conformal Classification to Enhance AI Trustworthiness and Reduce Uncertainty in Medical Imaging - Forward Pathway, accessed on May 5, 2025, https://www.forwardpathway.us/mit-combines-test-time-augmentation-and-conformal-classification-to-enhance-ai-trustworthiness-and-reduce-uncertainty-in-medical-imaging
54. Better Aggregation for Test-Time Augmentation - CAML, accessed on May 5, 2025, https://caml.csail.mit.edu/2021/11/10/aggregation-test-time-augmentation/
55. How MIT Researchers Made AI Models More Trustworthy For High-Stakes Medical Imaging, accessed on May 5, 2025, https://quantumzeitgeist.com/how-mit-researchers-made-ai-models-more-trustworthy-for-high-stakes-medical-imaging/
56. Understanding and Improving Robustness and Uncertainty Estimation in Deep Learning - Publikationen der UdS - Universität des Saarlandes, accessed on May 5, 2025, https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/33949/1/Thesis_Sulb_David_Stutz_Sep2022.pdf
57. Mining Unstructured Medical Texts With Conformal Active Learning - arXiv, accessed on May 5, 2025, https://arxiv.org/html/2502.04372v1
58. (PDF) Conditional Conformal Risk Adaptation - ResearchGate, accessed on May 5, 2025, https://www.researchgate.net/publication/390671728_Conditional_Conformal_Risk_Adaptation
59. Bias Mitigation Through Conditional Conformal Prediction - Restack, accessed on May 5, 2025, https://www.restack.io/p/bias-mitigation-answer-conditional-conformal-prediction-cat-ai
60. Improving Uncertainty Quantification in Regression Problems through Conformal Training, accessed on May 5, 2025, https://www.mlmi.eng.cam.ac.uk/files/2022_-_2023_dissertations/improving_uncertainty_quantification_in_regression_problems.pdf
61. Time-Series Deep Learning and Conformal Prediction for Improved Sepsis Diagnosis in Non-ICU Hospitalized Patients - ResearchGate, accessed on May 5, 2025, https://www.researchgate.net/publication/386072506_Time-Series_Deep_Learning_and_Conformal_Prediction_for_Improved_Sepsis_Diagnosis_in_Non-ICU_Hospitalized_Patients
62. NeurIPS Poster An Information Theoretic Perspective on Conformal Prediction, accessed on May 5, 2025, https://neurips.cc/virtual/2024/poster/94151
63. Making the Improbable Possible: Generalizing Models Designed for a Syndrome-Based, Heterogeneous Patient Landscape - PubMed Central, accessed on May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10758922/