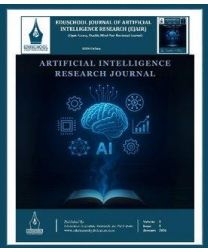
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Conformal Prediction for Interpretable Uncertainty in High-Stakes Decision Making

T Ramaprabha

Associate Professor, Department of Computer Science, Nehru Arts and Science College, Coimbatore, India

Article information

Received: 29th July 2026
Received in revised form: 31st August 2026
Accepted: 1st October 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0009

Abstract

Deep networks used in medicine, finance and autonomous systems report confidence scores that are often miscalibrated and that carry no guarantee once the data depart from the training distribution. Conformal prediction replaces the single score with a set of candidate outcomes and attaches a finite-sample guarantee to it. Under exchangeability, the set contains the true outcome with probability at least $1 - \alpha$, for any model and any data distribution. This article reviews the method as a tool for high-stakes decisions. It sets out the split conformal procedure and its coverage bounds, computes how the guarantee behaves at the calibration sizes available in clinical and safety-critical settings, and surveys the nonconformity scores used for classification and regression. It then examines three points at which the basic guarantee falls short in practice, namely conditional coverage across patients or subgroups, distribution shift, and losses other than miscoverage, and reviews the extensions that address each. Applications in medical imaging, pathology and language generation are assessed with the evidence on human use of prediction sets. The article closes with a proposed reporting protocol that ties the choice of α, the calibration data and the behaviour of set size to the decision the set is meant to support. The protocol has not been applied to a deployed system and is offered for testing.

Keywords:- Conformal Prediction, Uncertainty Quantification, Prediction Sets, Calibration, Distribution Shift, Risk Control, High-Stakes Decision Making

I. INTRODUCTION

A model that classifies a chest radiograph, flags a fraudulent transaction or proposes a drug dose produces a number alongside its answer, and that number is usually read as the probability that the answer is right. For modern deep networks this reading is unsafe. Guo et al. found that modern deep networks are systematically overconfident [1]. Their benchmarks covered image and text classification. A single temperature parameter fitted after training removed much of the gap on held-out data from the same distribution. Later work found that calibration depends strongly on architecture [2]. Popular recalibration methods are less calibrated than reported, and standard estimators cannot tell by how much [3]. Post-hoc calibration also falls short under dataset shift [4]. Calibration measured on a test set is an average property of the score, and it gives no statement about the error rate to expect on the cases a clinician sees next week.

Conformal prediction takes a different route [5], [6]. Instead of a point prediction with a score, it returns a set of labels or an interval, built so that the set contains the true outcome with a probability fixed in advance by the user. The guarantee holds in finite samples and for any underlying model, provided the calibration data and the new case are exchangeable. Two properties make the method attractive in high-stakes settings. The guarantee

does not require the model to be correct, so a poor model tends to yield larger sets on average rather than false confidence. The size of the set is also a readable statement of uncertainty, because a set with one label, a set with four labels and an empty set each carry a meaning on which a decision maker can act [7].

These properties are widely promoted, while the limits that accompany them receive less attention in applied work. Marginal coverage averaged over a population can coexist with poor coverage for a subgroup [8]. Exchangeability fails under the shifts that deployment brings [9]. The quantity a hospital cares about may also be a false negative rate rather than miscoverage [10]. This article reviews the method with these limits in view. Section II covers confidence scores and calibration. Section III states the split conformal procedure and its finite-sample behaviour. Section IV covers conditional coverage, shift and risk control. Section V reviews applications and human use of prediction sets, Section VI proposes a reporting protocol, and Section VII concludes.

II. CONFIDENCE SCORES WITHOUT GUARANTEES

A classifier with softmax output is calibrated when, among all inputs that receive confidence p for the predicted class, a fraction p is classified correctly. The property is defined over the input distribution. It is usually estimated by binning predictions and comparing average confidence with accuracy in each bin, which gives the expected calibration error used by Guo et al. [1]. Temperature scaling divides the logits by a scalar T fitted on a validation set. The ranking of classes is unchanged. A value of T above 1 softens the probabilities and lowers overconfidence. The method is simple and works well on data from the training distribution. It has become the default post-hoc correction.

Its value for decisions is limited in three ways, beginning with estimation. Kumar et al. showed that Platt scaling and temperature scaling are less calibrated than binned estimates suggest, and that the usual estimators cannot measure the remaining miscalibration, so a reported error near zero offers little assurance [3]. Dataset shift is a separate problem. Ovadia et al. compared uncertainty methods in a large benchmark under dataset shift [4]. In their results, post-hoc calibration on held-out data did not survive the shift. Methods that average over several models, such as ensembles of independently trained networks [11], gave stronger results than single networks with temperature scaling or Monte Carlo dropout [12]. Calibration also concerns averages over inputs that share a score, and it does not translate into a bound on the error rate of a specific decision rule, such as acting on the top label whenever confidence exceeds 0.9.

Bayesian approximations and ensembles give richer estimates of uncertainty. Their credible regions, however, are valid only if the posterior approximation is accurate, which is hard to check for networks with millions of parameters. High-stakes use needs a weaker and more practical statement, namely that a stated procedure misses the true outcome in at most a fixed fraction of cases, together with an output whose size carries the uncertainty. Rudin has argued that high-stakes decisions call for models whose reasoning can be inspected [13]. Kompa et al. argued that clinical models should quantify and communicate uncertainty and should be able to abstain on uncertain cases [14]. Conformal prediction meets the coverage requirement directly and supplies a set as its output. The set does not, however, explain the reasoning of the underlying model.

III. SPLIT CONFORMAL PREDICTION IN FINITE SAMPLES

A. The procedure and its guarantee

The split, or inductive, form of conformal prediction separates model fitting from calibration [15], [16]. A model is trained on one portion of the data. A nonconformity score s(x, y) is defined that grows as a candidate label y fits the input x less well. A common choice for classification is $s(x, y) = 1 - f_y(x)$, one minus the softmax probability of label y [17]. The score is computed on n held-out calibration pairs, giving $s_1, \ldots, s_n$. The threshold q is the $\lceil (n + 1)(1 - \alpha) \rceil$-th smallest of these scores, and the prediction set for a new input x is $C(x) = \{y: s(x, y) \leq q\}$. Any trained classifier can supply the score, and the model itself is not retrained.

If the calibration pairs and the test pair are exchangeable, the rank of the test score among all n + 1 scores is uniformly distributed. The set then satisfies:

$$1 - \alpha \leq P(Y \in C(X)) \leq 1 - \alpha + \frac{1}{n+1} \quad (1)$$

where the upper bound requires continuous scores without ties [7], [5]. The guarantee is marginal. It averages over the draw of the calibration set and the test point. It makes no claim about any particular input. Fig. 1 plots the empirical quantile level $\frac{\lceil (n + 1)(1 - \alpha) \rceil}{n}$ that the threshold must reach, for $\alpha = 0.05$, 0.1 and 0.2. The level exceeds $1 - \alpha$ by a margin that shrinks as n grows. At $n = 20$ and $\alpha = 0.1$ it equals 0.95. The threshold is finite only when $n \geq \frac{1}{\alpha} - 1$, so at least 19 calibration points are needed for $\alpha = 0.05$ and at least 9 for $\alpha = 0.1$.

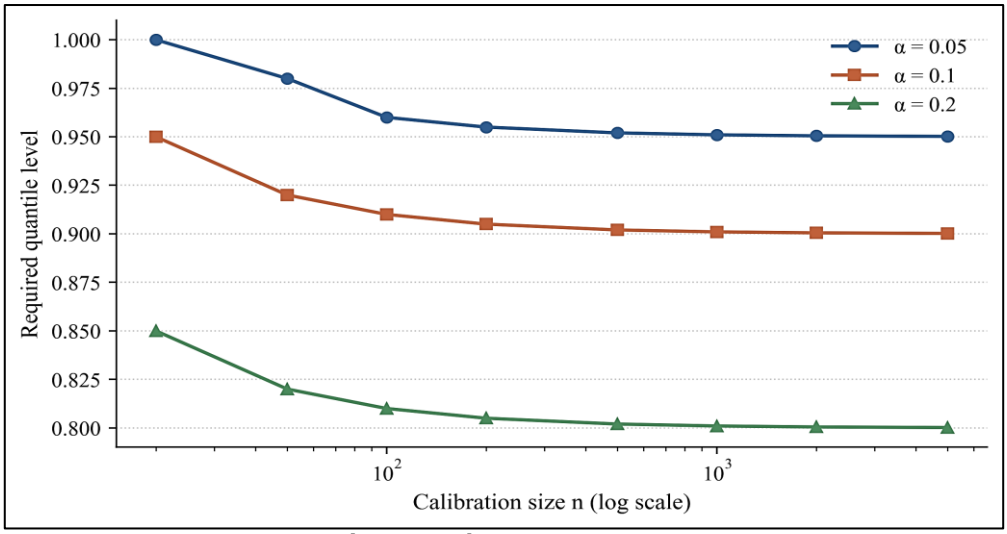


Fig. 1: Finite-sample quantile level $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ at which split conformal prediction sets its threshold, computed from the formula in Section III for calibration sizes n from 20 to 5000 and three miscoverage levels α. The level approaches the nominal $1 - \alpha$ as n grows.

B. Coverage given a fixed calibration set

A second distinction matters more for deployment. A hospital calibrates once and then applies the same threshold for months. The relevant quantity is then the coverage conditional on the particular calibration set that was drawn. For i.i.d. data and continuous scores, this conditional coverage follows a Beta distribution with parameters $n + 1 - l$ and l, where $l = \lfloor (n + 1)\alpha \rfloor$ [18], [7]. Its mean is close to $1 - \alpha$, but its spread is wide at small n. Fig. 2 plots the 5th, 50th and 95th percentiles of this distribution for $\alpha = 0.1$. The median stays close to 0.90 throughout. With $n = 100$, the 5th percentile is 0.848 and the probability that the realised coverage falls below 0.88 is 0.23, whereas with $n = 1000$ the 5th percentile rises to 0.884 and that probability falls to 0.02. Calibration sets of a few hundred cases are not unusual when labels need expert review, and at that size the coverage actually delivered can differ from the nominal level by several percentage points in either direction.

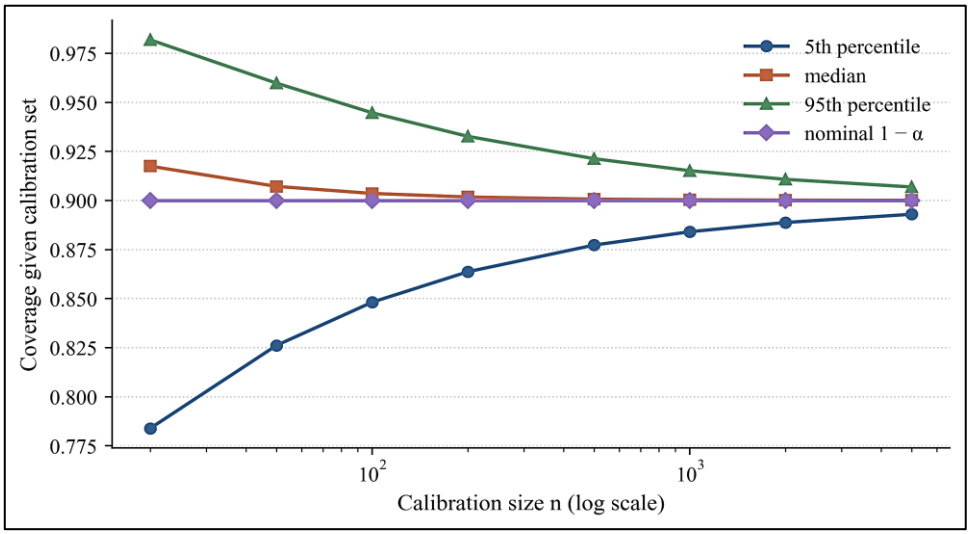


Fig. 2: Coverage conditional on the calibration set for split conformal prediction at $\alpha = 0.1$, computed from the Beta($n + 1 - l$, l) law with $l = \lfloor (n + 1)\alpha \rfloor$ stated in Section III [18], [7]. Lines show the 5th, 50th and 95th percentiles for n from 20 to 5000, with the nominal level for reference.

C. The choice of score

Any score yields valid sets. The score decides only how useful the sets are. Table 1 lists six of them. The threshold score of Sadinle et al. gives the smallest expected set size among procedures with the same coverage when the estimated class probabilities are correct [17]. In practice it tends to under-cover hard inputs and over-cover easy ones [7], [19]. Adaptive prediction sets (APS) add sorted class probabilities until the true label is reached, which gives larger sets on ambiguous inputs and better approximate conditional coverage [19]. The regularised variant RAPS penalises the long tail of unlikely classes, and on ImageNet its authors report sets often

5 to 10 times smaller than a Platt scaling baseline while achieving the target coverage [20]. For regression, conformalized quantile regression (CQR) calibrates the band between two fitted quantiles, so that interval width varies with the input [21], whereas the absolute residual score gives intervals of constant width [16].

Table 1. Nonconformity scores used in conformal prediction

Score	Task	Definition	Behaviour of the sets	Source
Threshold	Classification	One minus the softmax probability of the label	Smallest average size, uneven coverage across difficulty	[17]
APS	Classification	Cumulative sorted probability up to the label, randomised	Larger sets on ambiguous inputs	[19]
RAPS	Classification	APS with a penalty on labels beyond rank k	Smaller and more stable sets	[20]
Mondrian	Classification	Any score, calibrated within each class or group	Coverage within every category, fewer data per category	[5], [22]
Absolute residual	Regression	$\|y - \mu(x)\|$ for a fitted mean μ	Intervals of constant width	[16]
CQR	Regression	Distance outside a fitted quantile band	Width adapts to the input	[21]

IV. CONDITIONAL COVERAGE, SHIFT AND RISK CONTROL

A. Conditional and group coverage

Marginal coverage permits a procedure to cover 99% of easy cases and 70% of hard ones while averaging 90%, when about two thirds of cases are easy. In clinical use the hard cases are often those with the greatest consequences. Exact coverage conditional on every input value cannot be obtained without assumptions on the distribution [18]. Vovk analysed which forms of conditional validity inductive conformal predictors can and cannot achieve [18]. Foygel Barber et al. state the result that exact distribution-free conditional coverage forces intervals of infinite expected length at almost every point that is not an atom of the feature distribution [8]. They then examined weaker targets that remain achievable.

Coverage within groups is attainable when the groups are defined in advance. The Mondrian construction calibrates separately within each category, such as each class label or each patient subgroup, and gives the guarantee within every category at the cost of smaller calibration sets [5]. Romano et al. used this idea to obtain equalized coverage across protected groups [23]. Lu et al. interviewed radiologists about possible uses of conformal predictors, evaluated several conformal methods for skin lesion classification and adapted them to give equal coverage across groups of patient skin tone [24]. When the number of classes is large, calibration within each class runs short of data [22]. Ding et al. proposed clustering classes with similar score distributions so that each cluster has enough calibration examples [22]. The Beta law of Fig. 2 applies within each group, so a subgroup with 100 calibration cases inherits the wide spread computed for n = 100.

B. Distribution shift

Exchangeability fails when the deployed population differs from the calibration population, which is the normal condition once a model leaves the site where it was built. Tibshirani et al. showed that under covariate shift, where the distribution of inputs changes but the conditional distribution of outcomes does not, coverage is restored by weighting calibration scores with the likelihood ratio between the test and calibration input densities [9]. The ratio must be known or estimated, and errors in the estimate pass directly into the coverage. For data that arrive in sequence, adaptive conformal inference updates the working level after each case [25]. The update is:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t) \tag{2}$$

Here err_t is 1 when the set at time t missed the outcome and 0 otherwise. It needs no model of the shift. The long-run miscoverage frequency then converges to α for any sequence, with an error of order $\frac{1}{\gamma T}$ after T steps, although coverage in a short window can still depart from the target. Foygel Barber et al. extended the analysis to non-exchangeable data and bounded the loss of coverage by a weighted sum of total variation distances, which gives a quantitative account of how much coverage drift can cost [26].

C. Risk control

Many decisions depend on a loss other than miscoverage. A tumour segmentation system is judged by the fraction of tumour pixels it misses, and a multi-label screening tool by its false negative rate. Bates et al. introduced risk-controlling prediction sets, which choose a threshold so that the expected loss falls below a target with high probability over the calibration draw [27]. Conformal risk control gives the in-expectation guarantee $E[L(\lambda)] \leq \alpha$ for any loss bounded by B and monotone in a threshold λ [10]. The threshold is the smallest λ for which:

$$\frac{n}{n+1} R_n(\lambda) + \frac{B}{n+1} \leq \alpha \tag{3}$$

where R_n is the empirical risk on the calibration data. Both guarantees hold in finite samples. Learn then test treats the selection of any set of tuning parameters as a multiple testing problem [28]. This removes the monotonicity requirement and allows several risks to be controlled together [28]. These methods carry conformal reasoning from sets to thresholded decisions in general, which is the form most high-stakes systems take.

V. APPLICATIONS AND HUMAN USE OF PREDICTION SETS

Medical imaging has received the most attention. Angelopoulos et al. derived per-pixel uncertainty intervals for image-to-image regression with a user-specified confidence guarantee and evaluated them on quantitative phase microscopy, accelerated magnetic resonance imaging and electron microscopy [29]. Olsson et al. paired an AI system for the diagnosis and grading of prostate biopsies with conformal prediction [30]. The test set held 3059 biopsies from 676 men. They report that on test data from the training laboratory and scanner, 1 of 794 predictions for cancer diagnosis was incorrect with conformal prediction, against 14 errors without it. In the same setting, 175 predictions (22%) were flagged as unreliable. The same study reports that small external samples were enough to detect systematic differences in data from other scanners and laboratories, which points to a second use of conformal output as an alarm for the distribution shift discussed in Section IV.

Language generation is a newer target. Quach et al. sample candidate responses from a language model until a calibrated stopping rule is met, so that the returned set contains at least one acceptable response with a stated probability, and their tasks include radiology report generation [31]. Mohri and Hashimoto make generated answers less specific until the remaining claims are correct with a target probability [32]. They report correctness guarantees of 80% to 90% while most of the original output is retained [32]. In both cases the guarantee concerns a set or a reduced answer defined by a scoring function, and its value depends on how well that function separates acceptable from unacceptable outputs.

A coverage guarantee helps only if the people who receive the set use it well. Straitouri et al. designed a decision support system in which the expert must choose a label from the conformal set, and simulation experiments with synthetic and real expert predictions suggested that the system can help experts predict more accurately [33]. Cresswell et al. ran a pre-registered randomised controlled trial and found, with statistical significance, that participants given conformal sets were more accurate than those given fixed-size sets with the same coverage guarantee [34]. Both studies use controlled tasks. Evidence from clinicians making real decisions with conformal sets remains scarce.

Set size is the most interpretable output of the method. A singleton is a confident answer with a known error rate, a large set signals ambiguity, and an empty set under the threshold score signals that no label fits well, which can mark an input far from the training data. The set does not explain why the model is uncertain, so it complements the interpretable modelling advocated by Rudin rather than replacing it [13]. Multimodal systems offer a further use. A reliability-weighted fusion mechanism for emotion recognition proposed in this journal is designed to estimate how trustworthy each input stream is [35], and a conformal layer over its output would turn its output scores into sets with a stated coverage.

VI. A DECISION-ORIENTED REPORTING PROTOCOL

The work reviewed above suggests that a conformal guarantee is only as useful as the description that accompanies it. This article therefore proposes a reporting protocol for conformal systems in high-stakes deployment. It has six stages, and each stage yields one or two reported items. Fig. 3 shows them and Table 2 lists the items. The protocol is untested. Its items are derived from the failure points identified in Sections III and IV.

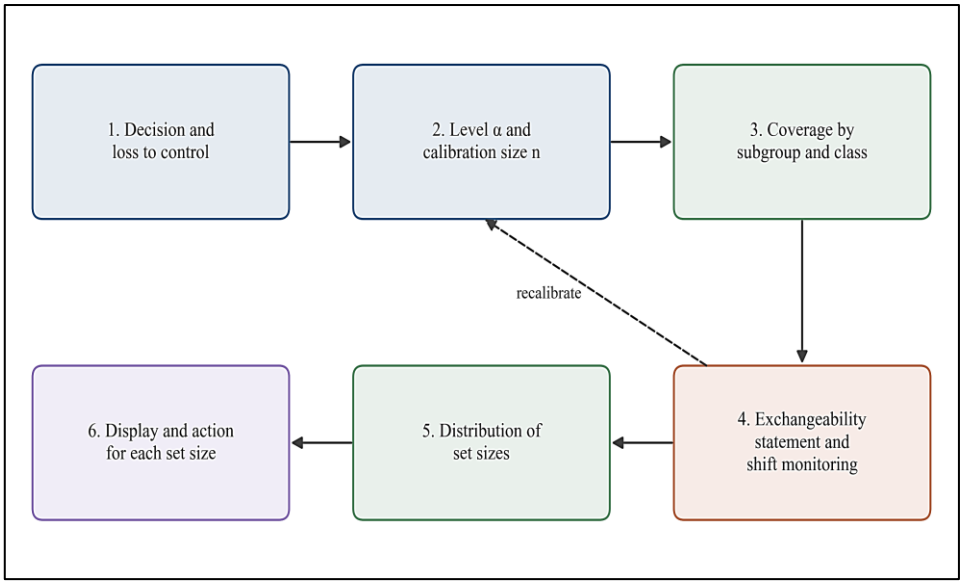


Fig. 3: Schematic of the proposed decision-oriented reporting protocol for conformal prediction in high-stakes deployment. Each stage produces a reported item listed in Table 2, and the monitoring stage feeds back to recalibration when the observed miscoverage departs from the target.

The protocol starts from the decision rather than from the model. The first stage names the action that the set supports, such as referral to a specialist whenever the set contains more than one diagnosis. It also names the loss attached to each kind of error, which determines whether miscoverage, a false negative rate or another risk should be controlled [10], [28]. The second stage fixes α with the people who bear the consequences and records the calibration size n together with the spread implied by the Beta law. A reader then sees, for example, that n = 200 at α = 0.1 places the 5th percentile of realised coverage at 0.864 (Fig. 2). The third stage reports coverage and set size for each predefined subgroup and class, with group sizes, since marginal coverage can hide failures in exactly these strata [8], [22].

The fourth stage states the exchangeability assumption in operational terms. It names the sites, devices and time period of the calibration data. It also specifies the monitoring rule that will detect shift, for example a running miscoverage rate with an adaptive update of the kind described in Section IV [25]. The fifth stage reports the distribution of set sizes rather than the average, because a mean size of 1.6 may consist of many singletons and a few very large sets. Singletons and very large sets call for different workflows. The sixth stage records how sets re shown to the user and what the user is expected to do with each size, since the benefit observed in a simulation study and a human trial depends on how the sets are built and offered [33], [34]. Table 2 pairs each item with the failure it is meant to prevent.

Table 2. Proposed reporting items for a deployed conformal system

Item	Content reported	Failure it guards against	Basis
Decision and loss	Action taken for each set outcome, loss controlled	Guarantee on the wrong quantity	[10], [28]
Level and calibration size	α, n and the 5th percentile of conditional coverage	Over-reading the nominal level	Fig. 2, [7]
Stratified coverage	Coverage and set size per subgroup and class, with counts	Hidden undercoverage in a subgroup	[23], [22]
Exchangeability statement	Sites, devices and period of the calibration data	Unnoticed shift	[9], [26]
Monitoring rule	Running miscoverage and the trigger for recalibration	Drift after deployment	[25]
Set-size distribution	Histogram of sizes, share of singletons and empty sets	Averages that hide large sets	[20]
Presentation	Display and expected user action for each size	Sets ignored or misused	[33], [34]

Three limitations apply. The protocol has not been applied to a deployed system, and its items may prove incomplete when tried in a specific clinical workflow. The values in Fig. 2 assume continuous scores and i.i.d. data, so they describe the best case rather than what a shifted deployment will achieve. The evidence on human use comes from a small number of controlled studies, and the claim that prediction sets improve expert decisions

in clinical practice is still a hypothesis. Validation would require a prospective study in which the reported items are compared with the coverage and decisions observed after deployment.

VII. CONCLUSION

Conformal prediction gives deep learning systems an output whose marginal error rate is set by the user and holds without assumptions on the model. The guarantee is marginal and depends on exchangeability. At calibration sizes of a few hundred, its realised value also spreads over several percentage points, as the Beta law in Fig. 2 shows. Extensions for group coverage, distribution shift and general risks address each of these limits in part, and each brings its own requirement for data or assumptions.

For high-stakes use, the method is most valuable when the set is tied to a named action and reported with its calibration size, its coverage by subgroup and its monitoring rule. The protocol in Table 2 is a proposal for that reporting. Testing it on real deployments, together with studies of how clinicians and other experts act on prediction sets, is the next step.

REFERENCES

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 70, 2017, pp. 1321–1330.

[2] M. Minderer *et al*., "Revisiting the calibration of modern neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 15682–15694.

[3] A. Kumar, P. Liang, and T. Ma, "Verified uncertainty calibration," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[4] Y. Ovadia *et al*., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 2nd ed. Cham, Switzerland: Springer, 2022.

[6] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, 2008.

[7] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.

[8] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Inf. Inference*, vol. 10, no. 2, pp. 455–482, 2021.

[9] R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[10] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[12] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 48, 2016, pp. 1050–1059.

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[14] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: Communicating uncertainty in medical machine learning," *npj Digit. Med.*, vol. 4, no. 1, Art. no. 4, 2021.

[15] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in *Proc. 13th Eur. Conf. Mach. Learn. (ECML)*, LNCS, vol. 2430, 2002, pp. 345–356.

[16] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Statist. Assoc.*, vol. 113, no. 523, pp. 1094–1111, 2018.

[17] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *J. Amer. Statist. Assoc.*, vol. 114, no. 525, pp. 223–234, 2019.

[18] V. Vovk, "Conditional validity of inductive conformal predictors," *Mach. Learn.*, vol. 92, nos. 2–3, pp. 349–376, 2013.

[19] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 3581–3591.

[20] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[21] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[22] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani, "Class-conditional conformal prediction with many classes," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 64555–64576.

[23] Y. Romano, R. Foygel Barber, C. Sabatti, and E. Candès, "With malice toward none: Assessing uncertainty via equalized coverage," *Harvard Data Sci. Rev.*, vol. 2, no. 2, 2020.

[24] C. Lu, A. Lemay, K. Chang, K. Höbel, and J. Kalpathy-Cramer, "Fair conformal predictors for applications in medical imaging," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, 2022, pp. 12008–12016.

[25] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 1660–1672.

[26] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *Ann. Statist.*, vol. 51, no. 2, pp. 816–845, 2023.
[27] S. Bates, A. N. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, "Distribution-free, risk-controlling prediction sets," *J. ACM*, vol. 68, no. 6, pp. 1–34, 2021.
[28] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *Ann. Appl. Statist.*, vol. 19, no. 2, pp. 1641–1662, 2025.
[29] A. N. Angelopoulos *et al*., "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 162, 2022, pp. 717–730.
[30] H. Olsson *et al*., "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nat. Commun.*, vol. 13, Art. no. 7761, 2022.
[31] V. Quach *et al*., "Conformal language modeling," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
[32] C. Mohri and T. Hashimoto, "Language models with conformal factuality guarantees," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 235, 2024, pp. 36029–36047.
[33] E. Straitouri, L. Wang, N. Okati, and M. Gomez Rodriguez, "Improving expert predictions with conformal prediction," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 202, 2023, pp. 32633–32653.
[34] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis, "Conformal prediction sets improve human decision making," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 235, 2024, pp. 9439–9457.
[35] N. Chandran, "Audio-visual-textual fusion for emotion recognition: A neurologically inspired attention mechanism," *Eduschool J. Artif. Intell. Res.*, vol. 1, no. 2, pp. 8–14, 2026, doi: 10.63090/EJAIR/3139.7336.0007.