Marginal Coverage Hides a Reduced-Ejection-Fraction Reliability Gap: Severity-Conditional Conformal Intervals on Public Echocardiography

Robert Sneiderman

Independent
robbysneiderman@gmail.com

Abstract. A conformal ejection-fraction (EF) interval that is calibrated on average can still be unreliable on the patients a cardiac-function screen exists to catch. Split-conformal prediction certifies *marginal* coverage, a guarantee a model can meet while systematically under-covering a clinical severity band. On the public EchoNet-Dynamic cohort, a distribution-free EF interval attains its 90% marginal target (empirical coverage 0.901 over 1277 evaluation exams), yet the clinically critical reduced-EF stratum (EF<40) is covered at only 0.694 (Wilson 95% CI $[0.618, 0.760]$, excluding the 0.90 target), a 0.206 shortfall, while the preserved stratum (EF≥50) is over-covered at 0.935: the certificate is least trustworthy exactly for the sickest patients. We frame this as a bias in the *reliability* of uncertainty, distinct from bias in point accuracy, and show it is a calibration artifact. Severity-conditional (Mondrian) recalibration on the true band restores reduced-EF coverage to 0.925 at the cost of width (56.6pp vs 27.3pp mean); deployed with predicted severity it recovers only 0.738, because the point model's own error near the 40-point boundary reassigns some reduced patients to narrower bands. That marginal coverage can hide subgroup under-coverage is established and not claimed as new; the contribution is where it lands on an EF screen, its per-band width cost, and the separation between conditioning on known versus predicted severity.

Keywords: Echocardiography · Ejection fraction · Conformal prediction · Coverage equity · Uncertainty calibration · Subgroup reliability.

1 Introduction

Automated ejection-fraction reading is accurate on average: a video model reaches an EF mean absolute error near four points on the public EchoNet-Dynamic benchmark [10]. Average accuracy is not the quantity a clinician acts on for one patient. A reduced EF below 40 routes a patient onto a different treatment and surveillance pathway, and in cardio-oncology serial EF gates cancer-therapy decisions [6]. A model trustworthy in aggregate but loose exactly on the reduced-EF patients fails where the decision is most consequential.

Conformal prediction [16,1] wraps any EF model in a finite-sample coverage guarantee under exchangeability alone, which suits a deployed screen. The guarantee is *marginal*: it controls coverage averaged over the cohort and says nothing about coverage inside a fixed subpopulation, so when the score distribution differs across groups a single pooled quantile over-covers the easy majority and under-covers the hard minority while the marginal number stays on target. For EF the hard minority is the reduced band, both the smallest group and the one a screen is built to protect.

Uncertainty fairness, not accuracy fairness. Two distinct fairness questions attach to a medical predictor. *Accuracy fairness* asks whether point error is equal across groups. *Uncertainty fairness*, the subject here, asks whether the guarantee attached to each prediction holds equally across them: whether the 90% interval covers 90% of the reduced-EF patients or only 90% on average. They can move together, and here they do, but a group can be accurate enough in aggregate yet receive a certificate that silently under-delivers. The axis is not unexamined [9,5], and it is the one a distribution-free certificate speaks to. On public EchoNet-Dynamic a split-conformal EF interval attains the 90% marginal target (0.901) while the reduced-EF stratum is under-covered at 0.694, a shortfall whose 95% interval excludes the target and exceeds its noise floor; conditioning on the true band restores it to 0.925, on the predicted band only to 0.738 (Table 1, Fig. 1).

Explicit non-claims. (i) This is not a new EF estimator and no point-accuracy gain is claimed; the base network is a re-trained baseline whose residuals feed calibration. (ii) Neither the conformal machinery nor the group-conditional repair is novel [15,12]; they are the instrument. (iii) The strata are EF-severity bands, a clinical proxy, not demographic groups. (iv) That marginal coverage can hold while a subgroup under-covers is established [8,4,5] and we claim no priority for it.

Terminology. Ejection fraction (EF) is the percentage of blood the left ventricle expels per beat. Below 40 is *reduced*, impaired systolic function that puts a patient on guideline heart-failure therapy and closer surveillance; 40 to 50 is *mildly reduced*; 50 or above is *preserved*, the normal range. Those bands are the strata throughout. *Coverage* is the fraction of patients whose true EF falls inside the reported interval, here targeted at 0.90, and a *certificate* is the coverage guarantee attached to an interval. *Marginal* coverage is that promise averaged over the cohort, *conditional* coverage the same promise inside a named subgroup. *Split-conformal calibration* sets one half-width from the error quantile of a held-out fold the model never trained on; *severity-conditional* (Mondrian) recalibration fits one half-width per band, so each band carries its own guarantee.

2 Related Work

Split conformal prediction and its marginal guarantee. Inductive (split) conformal prediction turns any point predictor into an interval predictor with a

finite-sample coverage guarantee under exchangeability alone [16,1]: an empirical quantile of held-out nonconformity scores fixes the interval half-width. The guarantee is marginal, silent about any fixed subpopulation, so a single marginal number can sit at nominal while one band runs well below it. Conformalized quantile regression [12] and adaptive prediction sets [13] vary width or set size with local uncertainty, but adaptivity follows wherever the fitted model already varies, so it does not equalize coverage across a partition that model never separated.

Conditional validity and distribution shift. The distance between marginal and conditional coverage is the subject of Mondrian, group-conditional conformal prediction, which partitions the calibration set by a declared taxonomy and computes one quantile per cell, recovering coverage within each cell at the price of cell-wise calibration sample [15]. Exact feature-conditional coverage is unattainable distribution-free, which is why one asks for coverage conditional on a coarse, pre-declared partition; the calibration studied here is a Mondrian taxonomy whose cells are clinical EF-severity bands. Coverage also drops when calibration and deployment distributions differ, which weighted conformal prediction corrects when the likelihood ratio is estimable [14]. That is not the mechanism here: both folds come from the same exchangeable pool, and the gap is within-distribution heterogeneity, a minority band with a heavier error tail averaged over by the pooled quantile.

Subgroup reliability of conformal coverage. That marginal coverage can hold while a subgroup under-covers is established, and we claim no priority for the observation. Mehrtens et al. [8] state it directly for medical image classification, reporting that a 90% marginal guarantee "does not guarantee this for individual instances or structured subgroups". Lu et al. [4] measure it in this venue's own setting, finding lower class-conditional coverage for the ambiguous middle grades of a stenosis-severity rating; a companion study [5] defines coverage disparity across Fitzpatrick skin type. Fairness of *uncertainty* rather than accuracy has been audited outside conformal prediction [9], and equalized coverage originates as a fairness target in regression [12]. What is new here is not the phenomenon but where it lands and what it costs: the under-covered band is the clinical extreme a screen exists to catch rather than an ambiguous middle grade, the setting is a real-valued interval rather than a class set, and we report the deployable repair's shortfall, its noise floor, and its interaction with abstention rather than a repair assumed to work.

Whether equal coverage is the right target. Cresswell et al. [2] argue the other way, that equalizing coverage is not automatically the right objective. We take the narrower position: under-coverage at a deployment site is a defect against the guarantee the certificate advertises, and we report the width a repair spends rather than assuming it is free.

Clinical setting, resolution floor, and alternative routes. EchoNet-Dynamic is a public, video-level benchmark with paired expert EF labels [10], and left-ventricular EF carries fixed clinical decision thresholds at which the care pathway changes [6], so the reduced band is where a silently under-covering interval is most consequential. Because per-stratum coverage is an empirical quantity, the minority band limits how small a real gap the study can resolve, a floor we take from the Dvoretzky–Kiefer–Wolfowitz inequality with the tight Massart constant [7]. Two alternatives bear on the result: the bin-conditional construction of Randahl et al. [11] combines per-bin intervals over a grid of candidate values without reading the test point's bin, and selective prediction abstains on hard cases [3]; Section 5 reports what each buys and gives up.

3 Method

Task and base predictor. Let X be an echocardiogram clip and $Y \in [0, 100]$ its expert ejection fraction. A base network $\hat{f}$ predicts EF from X, and the target is a prediction interval $C(X) \subseteq \mathbb{R}$ satisfying $\Pr[Y \in C(X)] \geq 1 - \alpha$ marginally and, beyond that, inside each clinical severity band. The level is fixed at $\alpha = 0.1$ throughout.

Split-conformal interval and severity strata. The nonconformity score is the absolute residual $s(x, y) = |y - \hat{f}(x)|$. Split-conformal calibration holds out a fold the model never trained on and takes the $\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil$-th smallest of its scores as the half-width $\hat{q}$, giving $C(x) = [\hat{f}(x) - \hat{q},\ \hat{f}(x) + \hat{q}]$ [16]. That single $\hat{q}$ meets the marginal guarantee and applies the same width to every band. The EF range is partitioned into three pre-declared clinical bands following standard thresholds [6]: $S_1 = \{Y < 40\}$ (reduced), $S_2 = \{40 \leq Y < 50\}$ (mildly reduced), and $S_3 = \{Y \geq 50\}$ (preserved), fixed before any calibration. It is the Mondrian taxonomy of the conditional calibration.

Algorithm 1 Severity-conditional (Mondrian) conformal EF intervals. TRUE reads the expert label, PRED the point prediction, on the calibration side and at test time alike; holding $\hat{q}_k = \hat{q}$ for every k recovers the marginal baseline.

Require: calibration fold $\{(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$, model $\hat{f}$, level α, variant $V \in \{\text{TRUE}, \text{PRED}\}$; band map $b(v) = 1, 2, 3$ for $v < 40$, $40 \leq v < 50$, $v \geq 50$; quantile $Q(S)$ = the $\min\big(\lceil (|S|+1)(1-\alpha) \rceil, |S|\big)$-th smallest element of S

$s_i \leftarrow |y_i - \hat{f}(x_i)|$ and $v_i \leftarrow y_i$ if $V = \text{TRUE}$, else $v_i \leftarrow \hat{f}(x_i)$ ▷ score, band source

$\hat{q} \leftarrow Q(\{s_1, \ldots, s_{n_{\text{cal}}}\})$ ▷ pooled half-width

for $k = 1, 2, 3$ **do**

  $A_k \leftarrow \{\, i : b(v_i) = k \,\}$, and $\hat{q}_k \leftarrow Q(\{s_i\}_{i \in A_k})$ if $A_k \neq \emptyset$, else $\hat{q}_k \leftarrow \hat{q}$

for a test exam x with expert label y: $k^\star \leftarrow b(y)$ if $V = \text{TRUE}$, else $k^\star \leftarrow b(\hat{f}(x))$

return $C(x) = \big[\hat{f}(x) - \hat{q}_{k^\star},\ \hat{f}(x) + \hat{q}_{k^\star}\big]$

Severity-conditional calibration and the two band rules. One quantile is fitted per band from that band's calibration scores only (Algorithm 1) [15], so each band carries its own finite-sample guarantee; a band too sparse to reach the level falls back to the pooled $\hat{q}$. Assigning an exam to a band needs a severity label, which is the reduced-versus-preserved decision the screen is trying to make, so two rules bound what recalibration can do. The *true-band* rule reads the ground-truth EF: not deployable, but it measures how much of the gap is a calibration artifact removable by conditioning. The *predicted-band* rule reads $\hat{f}(x)$: deployable, and what a fielded screen would run. Each rule applies on both sides, to the calibration exams forming a band and to the test exam assigned, so the difference between the rules isolates the part of the reduced-EF gap downstream of the point model's severity error.

Reported quantities and sample-size floor. For each calibration and band, the evaluation fold gives empirical coverage $\widehat{\mathrm{cov}}_k = \frac{1}{m_k}\sum_{i\in T_k}\mathbf{1}[Y_i \in C(X_i)]$ with test count $m_k = |T_k|$, mean width $\bar{w}_k$, and the worst-stratum gap $G = \max_k[(1-\alpha) - \widehat{\mathrm{cov}}_k]_+$. Coverage strata are cut on the true EF for every calibration, so the before-and-after comparison is on one fixed partition; the two variants differ only in how the calibration quantile is assigned, not in how coverage is audited. Coverage being an empirical proportion, the tight DKW inequality resolves it to about $\pm\big(2m_k\big)^{-1/2}\sqrt{\log(2/\delta)}$ for a band of test count m_k [7]; m_k is reported with every band so a worst-stratum gap is judged against the precision its sample supports.

4 Data and Setup

The data is the public EchoNet-Dynamic release of apical-4-chamber echocardiogram cine loops with paired expert LVEF labels [10]. The base network is a torchvision `r2plus1d_18` initialised from Kinetics-400 weights and fine-tuned for 6 epochs with Adam at 10^{-4}, batch size 8, on 32-frame clips sampled at stride 2 and resized to 112×112, under a Huber point loss with auxiliary pinball terms. Its only role is to supply per-exam EF residuals for calibration, so it is not tuned for benchmark-leading accuracy and its 6.6pp test MAE sits above the roughly 4pp of the published model. Evaluation uses the official test split unchanged ($n_{\mathrm{eval}} = 1277$), so no evaluation exam is seen at any earlier stage, and the calibration fold ($n_{\mathrm{cal}} = 1493$) is drawn patient-disjointly from the official training split, so no patient contributes to both. The target level is fixed at $\alpha = 0.1$ before any per-stratum analysis, the conventional level for an interval reported beside a screening decision, and no result below involves a search over α. The three strata partition the evaluation fold into 160 reduced (EF$<$40), 125 mildly reduced ($40 \le$ EF$<$50), and 992 preserved (EF$\ge$50) exams. Every reported coverage and width is computed on the evaluation fold; the Mondrian quantiles are fit out-of-sample on the calibration fold. The base model's EF mean absolute error is 6.6pp overall and heteroscedastic across severity (Table 1); that per-band error spread is the mechanism a single pooled quantile cannot absorb.

5 Results

Table 1. Per-EF-severity-stratum results on public EchoNet-Dynamic, $n_{\text{eval}} = 1277$, target $1 - \alpha = 0.90$, widths in EF points (pp). EF MAE is a property of the point model, so it is listed once; Wilson 95% intervals are given for each marginal stratum and for the reduced band under each repair. Marginal split-conformal meets the target in aggregate (0.901) but under-covers the reduced band by 0.206; Mondrian on the true band restores it to 0.925, on the predicted band only to 0.738.

Stratum	n	EF MAE (pp)	Cov.	worst gap	Mean width	Wilson 95% CI
Marginal split-conformal ($\hat{q} = 13.64$, uniform width 27.28)						
EF<40	160	10.8	0.694	**0.206**	27.28	[0.618, 0.760]
40–50	125	7.9	0.896		27.28	[0.830, 0.938]
EF≥50	992	5.8	0.935		27.28	[0.918, 0.949]
Marginal	1277	6.6	0.901		27.28	–
Mondrian, predicted severity band (deployable)						
EF<40	160	–	0.738	**0.162**	33.02	[0.664, 0.800]
40–50	125	–	0.880		28.96	–
EF≥50	992	–	0.931		25.40	–
Marginal	1277	–	0.902		26.70	–
Mondrian, true severity band (in-principle repair)						
EF<40	160	–	0.925		56.58	[0.873, 0.957]
40–50	125	–	0.896	**0.004**	27.02	–
EF≥50	992	–	0.899		24.32	–
Marginal	1277	–	0.902		28.63	–

The reduced-EF coverage gap. Table 1 is the result. The marginal interval sits at 0.901, on target, and a one-number audit would pass it. Split by severity, the same interval covers the preserved band at 0.935 and the reduced band at only 0.694, a 0.206 shortfall below the 0.90 target. The gap is not sampling noise: the reduced band's 95% interval $[0.618, 0.760]$ excludes 0.90, and the shortfall is nearly twice the Dvoretzky–Kiefer–Wolfowitz resolution floor of 0.107 that this band's 160 exams support at $\delta = 0.05$ [7]; the mildly reduced band, at 0.896, is within its noise floor. The pooled quantile buys the aggregate number by over-covering the 992 preserved exams and letting the 160 reduced exams pay: honest on the easy majority, optimistic on the hard minority.

Conditioning repairs it, with a width cost. Fitting one conformal quantile per severity band recovers per-band coverage. On the true band, reduced-EF coverage rises from 0.694 to 0.925 and the worst-stratum gap collapses from 0.206 to 0.004 (Table 1, top vs bottom block). The cost is width, borne where it should be: the reduced band's mean width grows from 27.28pp to 56.58pp while the

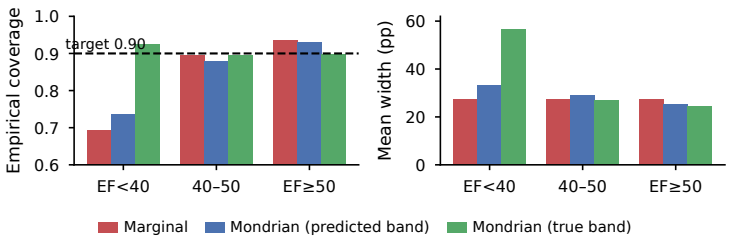

Fig. 1. Left: per-stratum empirical coverage for the three calibrations against the 0.90 target. Right: the price is width, concentrated on the reduced band, which needs a 56.6pp interval to reach target coverage.

preserved band tightens to 24.32pp, so recovered coverage is read against the width it spends rather than hidden inside an aggregate. The Wilson intervals in Table 1 separate the two repairs: the deployable 0.738 still excludes the 0.90 target, while under true-band conditioning every stratum contains it. The calibration fold behind these quantiles holds 188 reduced, 139 mildly reduced and 1166 preserved exams by true band.

Known severity versus predicted severity. The true-band result assigns each exam to its stratum using the label, unavailable at inference; the deployable version must assign from the predicted EF, which recovers reduced-EF coverage only to 0.738 (Table 1, middle block). The residual gap traces to the point model: near the 40-point boundary its 10.8pp reduced-band error reassigns some truly reduced patients to the preserved or mildly reduced band, where they receive that band's narrower quantile (12.2 and 16.2pp, against 18.0pp for the predicted-reduced band) and fall outside the interval. As Algorithm 1 makes explicit, the band rule applies on both sides, so prediction also moves calibration exams out of the reduced stratum (188 to 161) and the band whose quantile matters most is fitted on the fewest exams. The label is not, however, the only route to the repair: testing each *candidate* EF against the quantile of its own band never reads the label and attains the true-band coverage of Table 1 exactly, since an exam whose true EF lies in a band is covered precisely when its residual falls inside that band's quantile. What it gives up is contiguity, the resulting set being a union of grade ranges rather than one EF interval for most exams, which is why we report the banded form a screen can display.

The disparity survives abstention. One response to an untrustworthy interval is to abstain and route the exam to a human read [3]; abstention does not remove the disparity, it relocates it. At a half-width acceptance threshold of 17pp, the deployable policy answers 88.1% of exams automatically at 0.901 coverage on the answered set, but defers 69.4% of the truly reduced band against 1.8% of the preserved band. That cutoff is not tuned: the half-width takes only the three values quoted above, one per predicted band, so every cutoff in $[16.25, 18.02]$pp

is equivalent to the one reported and implements the rule "abstain on the band whose interval is widest". Nor is the disparity an artifact of that choice: at the only other non-degenerate cutoff the policy answers about three quarters of exams and defers a larger share of the reduced band still. Reduced patients are either under-covered automatically or disproportionately pushed to manual review, so a screen's audit must report per-stratum coverage and per-stratum deferral together, not either aggregate alone.

6 Discussion

A single marginal coverage number is not sufficient evidence that a conformal EF screen is reliable: here the 0.901 aggregate hides a 0.694 reduced-band coverage, a failure of the calibration rather than of the point model alone, which is why conditioning on severity repairs it in principle while the deployable repair depends on how well severity is predicted. A severity band is a conditioning variable, so this could be read as ordinary label-conditional coverage under a fairness name. What makes it a fairness finding is which band fails: the reduced-EF band is the clinically vulnerable minority a screen exists to catch, the smallest stratum (160 of 1277) and the one whose care pathway diverges at a fixed threshold [6]. The sickest patients hold the least trustworthy 90% interval while the marginal number reads as calibrated, assurance allocated in inverse proportion to clinical need. The takeaway is a procedure, not a model: audit a distribution-free certificate conditionally on the clinical strata it is trusted within, report each stratum's test count so a small band's gap is judged against its supported precision, and audit any abstention policy on the same partition. Reporting only marginal coverage, or only per-group accuracy, leaves the disparity invisible.

7 Limitations

The severity bands are a coarse, label-defined proxy for clinical risk, so equity here is across severity strata, not demographic groups; a demographic-conditional audit is the natural extension where labels exist. Every per-stratum figure is finite-sample, the reduced band's 160 exams setting the resolvable-gap floor (± 0.107 at $\delta = 0.05$). Coverage is distribution-free only under exchangeability of the two folds, so acquisition or site shift would need a shift correction [14]. The effect size is coupled to the backbone, trained once rather than under several seeds: a stronger model could show a smaller disparity and a narrower true-band/predicted-band gap, and the reported size carries no training-run variance. The disparity still follows from heteroscedastic per-stratum error, so a conditional audit stays necessary.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning (2023), arXiv:2107.07511
2. Cresswell, J.C., Kumar, B., Sui, Y., Belbahri, M.: Conformal prediction sets can cause disparate impact. In: International Conference on Learning Representations (ICLR) (2025), arXiv:2410.01888
3. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2017), arXiv:1705.08500
4. Lu, C., Angelopoulos, A.N., Pomerantz, S.: Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) (2022), arXiv:2207.02238
5. Lu, C., Lemay, A., Chang, K., Hoebel, K., Kalpathy-Cramer, J.: Fair conformal predictors for applications in medical imaging. In: AAAI Conference on Artificial Intelligence. pp. 12008–12016 (2022)
6. Lyon, A.R., López-Fernández, T., Couch, L.S., et al.: 2022 esc guidelines on cardio-oncology developed in collaboration with the european hematology association, the european society for therapeutic radiology and oncology and the international cardio-oncology society. European Heart Journal **43**(41), 4229–4361 (2022). https://doi.org/10.1093/eurheartj/ehac244
7. Massart, P.: The tight constant in the dvoretzky-kiefer-wolfowitz inequality. The Annals of Probability **18**(3), 1269–1283 (1990). https://doi.org/10.1214/aop/1176990746
8. Mehrtens, H., Bucher, T., Brinker, T.J.: Pitfalls of conformal predictions for medical image classification. arXiv preprint (2025), arXiv:2506.18162
9. Mehta, R., Shui, C., Arbel, T.: Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In: Medical Imaging with Deep Learning (MIDL), PMLR. vol. 227, pp. 1453–1492 (2024), arXiv:2303.03242
10. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**, 252–256 (2020). https://doi.org/10.1038/s41586-020-2145-8
11. Randahl, D., Williams, J.P., Hegre, H.: Bin-conditional conformal prediction of fatalities from armed conflict. arXiv preprint (2024), arXiv:2410.14507
12. Romano, Y., Patterson, E., Candès, E.J.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems (NeurIPS) (2019), arXiv:1905.03222
13. Romano, Y., Sesia, M., Candès, E.J.: Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems (NeurIPS) (2020), arXiv:2006.02544
14. Tibshirani, R.J., Foygel Barber, R., Candès, E.J., Ramdas, A.: Conformal prediction under covariate shift. In: Advances in Neural Information Processing Systems (NeurIPS) (2019), arXiv:1904.06019
15. Vovk, V.: Conditional validity of inductive conformal predictors. In: Proceedings of the Asian Conference on Machine Learning (ACML) (2012), arXiv:1209.2673
16. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005). https://doi.org/10.1007/b106715