

Risk stratification through class-conditional conformal estimation: A strategy that improves the rule-out performance of MACE in the prehospital setting

Juan José García

JJGARCIA@CS.UNC.EDU

*Department of Computer Science
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA*

Nikhil Sarin

NIKHILSA@EMAIL.UNC.EDU

*Department of Computer Science
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA*

Rebecca R. Kitzmiller

KITZM002@EMAIL.UNC.EDU

*School of Nursing
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA*

Ashok Krishnamurthy

ASHOK@RENCI.ORG

*Department of Computer Science
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA*

Jessica K. Zègre-Hemsey

JZHEMSEY@EMAIL.UNC.EDU

*School of Nursing
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA*

Abstract

Accurate risk stratification of clinical scores is important to mitigate adverse outcomes in patient care. In this study we explore whether class-conditional conformal estimation can yield better risk stratification cutoffs, as measured by rule-out and rule-in performance. In the binary setting, the cutoffs are chosen to theoretically bound the false positive rate (FPR) and the false negative rate (FNR). We showcase rule-out performance improvements for the task of 30-day major adverse cardiac event (MACE) prediction in the prehospital setting over standard of care HEART and HEAR algorithms. Further, we observe the theoretical bounds materialize 96% and 77% of the time for FPR and FNR respectively across multiple datasets. Improving risk score accuracy is important since inaccurate stratification can lead to significant negative patient outcomes. For instance, in the case of MACE prediction, better rule-out performance translates into less delay of time dependent therapies that restore bloodflow to the compromised myocardium, thereby reducing morbidity and mortality.

1. Introduction

Risk stratification consists of categorizing patients according to their risk, where each category is coupled with a care path. Accurate risk stratification is important to mitigate adverse patient outcome or inappropriate resource utilization. Using a risk score (i.e. sparse linear model with integer coefficients) is a commonplace strategy to measure risk in healthcare (Ustun and Rudin (2019)) and cutoffs are generally used to stratify the scores. Unfortunately, approaches that estimate these cutoffs either result in stratification errors (Fluss et al. (2005)); or require the specification of six, seldom available, variables (Tortorella (2000)).

In this work, we explore whether class-conditional conformal estimation (CC) (Vovk (2012); Lei (2014)) can estimate cutoffs that mitigate stratification errors with the specification of two variables instead of six. Intuitively, by leveraging a population sample, CC estimates two stratification cutoffs that theoretically upper bound the the false-negative-rate (FNR) and the false-positive-rate (FPR). The upper bounds are predetermined by the practitioner and act as a tolerance level that controls the maximum proportion of missed cases and false alarms. This in turn improves performance by reducing false negatives and false positives with a tradeoff of ambiguity, a strategy also known as selective classification (Cordella et al. (1995); El-Yaniv et al. (2010)). For instance, in Figure 1, over a test sample for the task of major adverse cardiac events (MACE) prediction, we observe that cutoffs estimated with CC satisfy a 10% upper bound constraint for both FPR and FNR; the region between cutoffs is the ambiguous or intermediate-risk category.

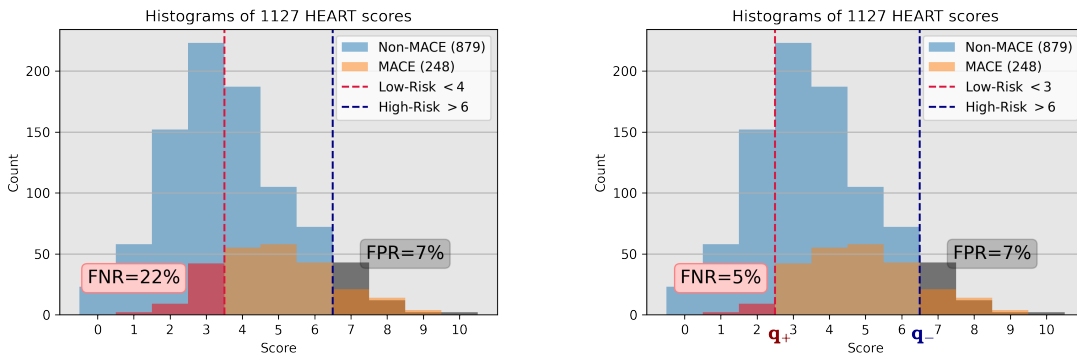


Figure 1: Blue histogram corresponds to risk scores for cases that developed MACE. Orange histogram corresponds to risk scores of cases that did not develop MACE. The vertical lines match the cutoffs used to risk stratify cases according to the literature (left) and estimated with CC (right). Highlighted in red are the false negative (FN) cases and in black the false positives (FP) cases.

We evaluate CC on the stratification of MACE risk scores in the prehospital setting, where the standard of care (HEART score) and proposed variants (HEAR score) lack sufficient rule-out performance to meet the needs of cardiologists (Cooper et al. (2023)). Improving these scoring tools holds the potential for earlier identification of patients in need of life saving definitive treatment, thus improving appropriate use of limited emergency

services resources. Furthermore, we explore the utility of using CC as a way to improve the AUROC performance of state-of-the-art ML score estimator FasterRisk (Liu et al. (2022)) on twelve benchmark datasets.

Our contributions are summarized as follows:

- We are the first study to evaluate the utility of class-conditional conformal estimation (CC) for risk stratification of clinical scores.
- We provide evidence that recalibrating the cutoffs to the deployment population using CC improves the rule-out performance of standard of care HEART score by 46% in sensitivity and 5% in negative-predictive-value (NPV) without exceeding the prespecified limit on the FPR. This recalibration also resulted in more intermediate risk cases (13%) (See section 5.2).

Generalizable Insights about Machine Learning in the Context of Healthcare

Our empirical evaluation of class conditional conformal estimation (CC) yields the following insights in the more general context of risk stratification of a binary event using scores:

- CC consistently improves the AUROC performance of estimated scores (e.g. FasterRisk) and pre-estimated scores (e.g. HEAR, HEART) without compromising their explainability, albeit with a tradeoff in coverage (See Sections 5.2 and 5.3).
- Across twelve benchmark datasets, the FPR and FNR theoretical bounds held 96% and 77% of the time respectively (See section 6.1). This is important when reassuring practitioners that cutoff estimation guarantees hold across different populations.
- Prevalence is important to inform a maximum sensible tolerance of FPR and FNR (See section 6.2). This may aid practitioners in choosing an initial FPR and FNR bounds for their task.
- A larger than tolerable proportion of ambiguous cases (i.e. cases categorized neither high-risk or low-risk scores) suggests the chosen risk score is not discriminative enough to sensibly meet FPR and FNR tolerance requirements in a given population. If this is the case, then we suggest to consider a different score (See section 6.3).

2. Related Work

Commonly, class-conditional conformal estimation is an algorithm to produce prediction intervals with conditional probabilistic guarantees (Vovk (2012)). It achieves this by compromising statistical efficiency (i.e. training data) in order to control the probability the true label belongs to the prediction interval. For our purposes, this methodology is particularly useful in that it is model-free (i.e. does not require assumptions on the predictor). Therefore, it is suitable to update risk scores without compromising their explainability. In the binary classification case, class-conditional conformal estimation may lead to three prediction intervals: $\{0\}$, $\{1\}$, $\{0, 1\}$, where the last interval implies ambiguity in the binary decision (Lei (2014)). Therefore, we use these prediction intervals to risk stratify a score,

where $\{0\}$ corresponds to low risk, $\{0, 1\}$ corresponds to intermediate risk, and $\{1\}$ corresponds to high risk. To the best of our knowledge, this is the first work that explores the utility of class-conditional conformal estimation in improving the risk stratification of scores used in healthcare.

On a similar note, group conditional conformal estimation has been used to guarantee prediction interval coverage over different groups. It has been tested on criminal scores like the Compass (Romano et al. (2020)) for the goal of algorithmic fairness. Our work explores class conditional conformal estimation, which differs in that it conditions on labels instead of features. Class conditional conformal estimation was explored in: Johansson et al. (2023) to control accuracy and PPV using a Modrian conformal algorithm (Vovk et al. (2005)); in Balasubramanian et al. (2009) to control accuracy on the task of Drug Eluting Stents complications; in Papadopoulos et al. (2017) to control accuracy given class imbalance on the task of stroke prediction. In contrast, we controlled the FNR and FPR for the stratification of risk scores using the split conditional conformal algorithm in Angelopoulos et al. (2023).

The task of MACE prediction in the prehospital setting is important because missed cases may delay time dependent therapies that mitigate irreversible damage to the compromised myocardium. Further, explainability is also important to aid decision making in the patient’s care (Elul et al. (2021)). Current common explainable scores to predict MACE are the HEAR and HEART scores. The only difference between the two is that the HEART score relies on a measurement of troponin (generally unavailable in the prehospital setting). Unfortunately, neither HEAR nor HEART achieve the ideal rule-out performance of MACE (i.e. 99% sensitivity and 99.5% negative predictive value (NPV) suggested by (Cooper et al. (2023))). Therefore, there is a need for explainable scores that bridge this performance gap. We conjecture conformal prediction can produce better MACE stratification cutoffs of HEAR(T) scores or further improve the performance of learned risk scores (e.g. FasterRisk).

Several works have addressed the way to estimate optimal cutoffs for selective classification (Tortorella (2000); Fluss et al. (2005); Geifman and El-Yaniv (2017)). Tortorella (2000) proposed an optimal reject rule for binary classifiers, which employs the Receiver Operating Characteristic (ROC) curve to strategically set cutoffs that maximize a utility function. This function is designed to weigh the benefits of correct classifications against the costs associated with errors. Despite its precision, this approach requires the prespecification of the utility function with costs that are not readily available. Similarly, Geifman and El-Yaniv (2017) introduced a methodology for determining cutoffs that allow deep neural networks to either classify or selectively reject instances based on desired risk levels. They employ confidence-rate functions, particularly, softmax response (SR) and Monte Carlo dropout (MC-dropout), to set thresholds that reliably maintain the desired error rates with high probability. While providing flexibility in managing risk, in the case of controlling FPR and FNR in binary classification their approach yields a weaker probabilistic guarantee compared to class-conditional conformal estimation. Lastly, concurrent work by Angelopoulos et al. (2024) suggests to control the maximum negative predictive value (NPV) and the maximum positive predictive value (PPV) using the learn-then-test algorithm proposed in Angelopoulos et al. (2021). In contrast, we explore controlling the FPR and FNR instead with the simpler split conditional conformal algorithm. An additional difference stems from measuring the utility of conformal estimation over risk scores due to their widespread use

to assess risk in healthcare (Ustun and Rudin (2019)), their explainable nature and their discrete values.

Risk scores (i.e. sparse linear models with integer coefficients) like FasterRisk are important in machine learning due to their explainability. Unfortunately, most ML methodologies used with various conditional conformal methods (e.g. convolutional neural networks (Angelopoulos et al. (2024); Lei (2014)), support vector machines (Balasubramanian et al. (2009)), ensembles of neural networks (Papadopoulos et al. (2017)), random forests (Johansson et al. (2023)), regression tree (MART) (Vovk (2012))) do not generally produce solutions that satisfy these requirements and are therefore limited in their practical applicability to aid clinical decision making (Elul et al. (2021)).

3. Methods

3.1. Class conditional conformal estimation (CC)

Class conditional conformal estimation has been proposed as a mechanism to obtain conditional coverage guarantees (Vovk (2012); Angelopoulos et al. (2023); Sadinle et al. (2019)). In the binary case, class conditional guarantees lead to control over the FPR and the FNR (Lei (2014)). Control over both these quantities is important in healthcare to limit the proportion of missed cases and false alarms. Furthermore, the limit is predetermined by the practitioner and thus it can be adjusted to meet the needs of the deployment site. More generally, let $\{X_j\}_{j \in I_+} \sim P(X|Y = +)$ correspond to a random sample of scores from positive cases indexed by I_+ . Similarly, let $\{X_j\}_{j \in I_-} \sim P(X|Y = -)$ correspond to a random sample of scores from negative cases indexed by I_- . From the class conditional conformal estimation algorithm in Angelopoulos et al. (2023) we can derive the cutoffs to be:

$$q_+ = -\text{Quantile}\left(\{-X_j\}_{j \in I_+}, \frac{\lceil (|I_+| + 1)(1 - \alpha_+) \rceil}{|I_+|}\right) \quad (1)$$

$$q_- = \text{Quantile}\left(\{X_j\}_{j \in I_-}, \frac{\lceil (|I_-| + 1)(1 - \alpha_-) \rceil}{|I_-|}\right) \quad (2)$$

Where q_+ and q_- are random variables. Let $n = |I_-| + |I_+|$ and consider a new score X_{n+1} . For risk stratification purposes we categorize $X_{n+1} < q_+$ as low-risk, $X_{n+1} > q_-$ as high-risk and $X_{n+1} \in [q_+, q_-]$ as intermediate risk. Since q_+ and q_- are estimated from a random sample (i.e. $X_{1:n}$), exchangeable w.r.t. X_{n+1} , then $\text{FNR} = P(X_{n+1} < q_+ | Y_{n+1} = +) \leq \alpha_+$ and $\text{FPR} = P(X_{n+1} > q_- | Y_{n+1} = -) \leq \alpha_-$. The last statement we formally refer to as proposition 1 and prove below. To aid the proof, we re-state lemma 1 from Tibshirani et al. (2019)

Lemma 1 (Tibshirani et al. (2019)): *If X_1, \dots, X_{n+1} are exchangeable random variables, then for any $\alpha \in (0, 1)$ we have:*

$$P(X_{n+1} \leq \text{Quantile}(\{X_{1:n}\} \cup \{\infty\}, \alpha)) \geq \alpha \quad (3)$$

Proposition 1 *1 Let q_+ and q_- be defined by equations 1 and 2 respectively. Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_{n+1}, Y_{n+1})$ be i.i.d. samples in $\mathbb{R} \times \{-, +\}$. Let $\alpha_+ \in \left[\frac{1}{|I_+|+1}, \frac{|I_+|}{|I_+|+1}\right]$ and $\alpha_- \in \left[\frac{1}{|I_-|+1}, \frac{|I_-|}{|I_-|+1}\right]$ then:*

$$P(X_{n+1} < q_+ | Y_{n+1} = +) \leq \alpha_+ \quad (4)$$

$$P(X_{n+1} > q_- | Y_{n+1} = -) \leq \alpha_- \quad (5)$$

Proof of Proposition 1.

Let $q'_+ = \text{Quantile}(\{-X_j\}_{j \in I_+} \cup \{\infty\}, 1 - \alpha_+)$. Note q'_+ is a random variable, $q_+ = -q'_+$ and $q'_+ \leq 0$. Also note $\{X_j\}_{j \in I_+} \sim P(X|Y = +)$ and assume $X_{n+1} \sim P(X|Y = +)$, by Lemma 1 it follows $P(-X_{n+1} \leq q'_+ | Y = +) \geq 1 - \alpha_+$ and thus $P(X_{n+1} < -q'_+ | Y = +) \leq \alpha_+$. Since $q_+ = -q'_+$ it follows $P(X_{n+1} < q_+ | Y = +) \leq \alpha_+$. Similarly, observe $q_- = \text{Quantile}(\{X_j\}_{j \in I_-} \cup \{\infty\}, 1 - \alpha_-)$. Note q_- is a random variable and $q_- \geq 0$. Also note $\{X_j\}_{j \in I_-} \sim P(X|Y = -)$ and assume $X_{n+1} \sim P(X|Y = -)$, by Lemma 1 it follows $P(X_{n+1} \leq q_- | Y = -) \geq 1 - \alpha_-$ and thus $P(X_{n+1} > q_- | Y = -) \leq \alpha_-$. ■


Intuitively, the quantile function in equation (2) estimates the $(1 - \alpha_-)$ largest score (i.e. q_-) amongst negative cases (i.e. $X_{i \in I_-}$). Assuming $Y_{n+1} = -$, and noting $X_{n+1}, X_{i \in I_-}$ are exchangeable, the proportion of times $X_{n+1} > q_-$ is at most α_- . In other words, the proportion of times we erroneously label a negative sample as high-risk is at most α_- (i.e. $P(X_{n+1} > q_- | Y_{n+1} = -) \leq \alpha_-$). Equations (1) and (4) have a similar intuition.

3.2. Risk scores for MACE prediction: HEAR and HEART

The HEAR score (Stopyra et al. (2018)) acronym represents its four components: (H)istory, (E)CG, (A)ge, (R)isk factors. HEART, an expanded version of HEAR, adds (T)roponin, a covariate generally unavailable in the prehospital setting. Each component is assigned a score between 0 and 2 to denote increasing risk levels providing a total score between 0 and 8 for the HEAR (See Figure 2) and a total score between 0 and 10 for the HEART. Our selection of ten covariates for calculating the HEAR score emerges from the intersection of features accessible to us with those used within the work of Stopyra et al. (2018). Specifically, in the History component of the HEAR score methodology, we evaluate based on three covariates: chest pain, nausea and/or vomiting, and diaphoresis. The scoring is straightforward: a score of 0 is assigned if none of these covariates are present. A score of 1 is given for the presence of one covariate. For the presence of two or more covariates, a score of 2 is allocated. The ECG component focuses on two covariates: significant ST depression, which scores a 2, and non-specific ST-T wave abnormalities, which earn a score of 1. A score of 0 is given when neither covariate is present. Age directly influences the scoring, with individuals aged 65 and over classified as a score of 2, those between 45 and 64 with a score of 1, and individuals under 45 with a score of 0. Finally, the Risk factors component evaluates the presence of hypercholesterolemia, hypertension, current smoker status, and diabetes. More than two present covariates result in a score of 2, one or two covariates present lead to a score of 1, and the absence of all four covariates indicates a score of 0. The HEART score follows a similar methodology and includes the evaluation of troponin levels as an additional factor. Troponin levels greater than three times the normal limit (i.e. 0.03 (ng/ml)) score a 2. Levels greater than one but less than or equal to three times the normal limit score a 1, and levels less than or equal to the normal limit score a 0.

3.3. Risk score estimator: FasterRisk

Due to the availability of data, it is sensible to use some of it to estimate a risk score. Accordingly, we chose FasterRisk (Liu et al. (2022)) as a score estimation algorithm to explore how to make use of our data: Do we use it for risk score estimation, or save some for cutoff estimation? To the best of our knowledge, FasterRisk is comparable to state-of-the-art score estimators in terms of AUROC and is faster in terms of runtime. We emphasize, the estimated risk scores can be chosen so as to satisfy the definition of interpretability proposed by Elul et al. (2021). This is harder to do for uninterpretable blackbox predictors (e.g. SVM, Neural Network).



(H)istory	Value
Highly suspicious	2
Moderately suspicious	1
Slightly suspicious	0
(E)CG	
Significant ST-Depression	2
Non-specific Repolarization	1
Normal	0
(A)ge	
> 64	2
45-64	1
< 45	0
(R)isk factors	
> 2	2
1-2	1
< 1	0
Total	0-8
If Total < 4	Low Risk
If Total > 6	High Risk

(H)istory	Value
Highly suspicious	2
Moderately suspicious	1
Slightly suspicious	0
(E)CG	
Significant ST-Depression	2
Non-specific Repolarization	1
Normal	0
(A)ge	
> 64	2
45-64	1
< 45	0
(R)isk factors	
> 2	2
1-2	1
< 1	0
Total	0-8
If Total < q_+	Low Risk
If Total > q_-	High Risk

Figure 2: (Left) Original HEAR score. (Right) HEAR+CC score. Note the difference is in the cutoffs. HEAR+CC cutoffs are tailored using a population sample.

4. Cohort

Data were collected by Zègre-Hemsey et al. (2021) in Mecklenburg county, North Carolina. Patients meeting the following criteria were included: over 21 years old, transported by ambulance to the ED with non-traumatic chest pain and/or anginal equivalents, acquired ≥ 1 ECG, without a diagnosis of STEMI. Emergency healthcare personnel collected clinical information in the ambulance (i.e. Prehospital setting). The primary outcomes recorded any MACE event (i.e. the acute manifestation of coronary heart disease and include non-ST elevation myocardial infarction (NSTEMI), and unstable angina (UA)). The observed prevalence was MACE (20%). These events occurred within 30 days post ED admission.

4.1. Cohort Selection

We divide the dataset (n=2883) into two cohorts: An internal cohort (n=1756 cases before 04/2016) for training and validation, and an external cohort (n=1127 cases after 04/2016) for testing as it is done in Takeda et al. (2022); Al-Zaiti et al. (2020).

Characteristic	Type	Internal(n=1756)	External(n=1127)
Age	Numerical	61.04 ± 30.96	59.97 ± 30.41
Gender(male)	Binary	936(53%)	629(55%)
Medical History	Type	Internal(n=1756)	External(n=1127)
Hypercholesterolemia	Binary	693(39%)	485(43%)
Hypertension	Binary	943(53%)	803(71%)
Current Smoker	Binary	368(20%)	283(25%)
Diabetes	Binary	509(28%)	354(31%)
Prior MI	Binary	303(17%)	245(21%)
Angina	Binary	42(2%)	80(7%)
Prior CABG	Binary	166(9%)	180(15%)
Prior PCI	Binary	124(7%)	6(0%)
CAD	Binary	349(19%)	271(24%)
Family History of CV Disease	Binary	204(11%)	81(7%)
Other	Binary	1753(99%)	1124(99%)
Symptoms	Type	Internal(n=1756)	External(n=1127)
Chestpain	Binary	992(56%)	644(57%)
Syncope	Binary	103(5%)	69(6%)
Shortness of Breath	Binary	417(23%)	282(25%)
Diaphoresis	Binary	114(6%)	89(7%)
Nausea and/or Vomiting	Binary	164(9%)	113(10%)
Palpitations	Binary	226(12%)	164(14%)
Other Symptoms	Binary	873(49%)	618(54%)
ECG Interpretation	Type	Internal(n=1756)	External(n=1127)
ST Elevation	$\{0, 1\}^{11}$	329.0(18%)	170.0(15%)
ST Depression	$\{0, 1\}^{11}$	500.0(28%)	217.0(19%)
T Wave Inversion	$\{0, 1\}^{11}$	558.0(31%)	270.0(23%)
Non-specific ST-T Wave Abnormalities	$\{0, 1\}^{11}$	252.0(14%)	180.0(15%)

Table 1: Statistics of covariates used as input to the FasterRisk model. Statistics are calculated separately for the internal and external cohorts. For the ECG interpretations, type $\{0, 1\}^{11}$ indicates a binary vector. The position corresponds to the ECG lead used for the interpretation.

4.2. Data Extraction

We select 23 covariates (see Table 1) commonly associated with MACE (Backus et al. (2013)) and available in the prehospital setting (Stopyra et al. (2018)). We discarded patients with a missing initial troponin value (25 total) or without an ECG date; less than 2% of patients had missing covariates imputed with a constant as is suggested by Le Morvan et al. (2021).

5. Results on Real Data

5.1. Evaluation Approach/Study Design

The goal is to assess the rule-in and rule-out performance of six algorithms on the 30-day prediction of MACE in the prehospital setting. The scores include: FasterRisk, HEART,

HEAR, FasterRisk+CC, HEART+CC and HEAR+CC. CC corresponds to a re-estimation of cutoffs using class conditional conformal estimation. We consider MACE as the 30-day outcome of death, re-infarction, new onset heart failure and others. Binary classification performance is measured in terms of: coverage, area under the reproducer-operator-curve (AUROC), accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV), sensitivity and specificity. Rule-out performance is measured in terms of sensitivity and NPV. Cutoffs are estimated with the internal cohort (i.e. 1700 cases that occurred before April 2016). In the case of FasterRisk, we estimate hyperparameters and cutoffs with 20% of the internal cohort and use the rest for estimating the model’s parameters. For hyperparameter estimation we used grid search to mitigate human-in-the-loop biases. Performance metrics were calculated over cases in the external cohort (i.e. 1100 cases that occurred after April 2016) with scores considered either high-risk or low-risk. We reported the prevalence of MACE within the selected group as well as the proportion of patients that were considered from the entire population (i.e. coverage or non-ambiguous).

We also benchmark the performance of adding CC to risk stratify the output of the FasterRisk model. The benchmark consists of a subset of three tabular datasets tested in Liu et al. (2022) and nine from Grinsztajn et al. (2022). For this experiment, we evaluate performance using Stratified 3-Fold cross-validation. For each fold, we further split the training set into a training subset (80%) and validation subset (20%). The validation subset is used to grid search the optimal hyperparameters and CC estimation. Like Liu et al. (2022), the primary metric we focused on to assess the model’s discriminatory power was the AUROC. We further measure prevalence and coverage to understand practical tradeoffs introduced by the extra cutoff. Code for the tabular benchmark is available online¹.

5.2. Results on MACE classification

Results in Table 2 suggest that adapting the cutoffs with CC improves rule-out performance in all cases, albeit with an increase in the proportion of cases labeled as intermediate risk (i.e. ambiguity). Given the prevalence of 20%, we conjecture the increase in performance comes from the dependence of NPV and sensitivity on false-negatives. Accordingly, controlling the proportion of false negatives increases both sensitivity by 64% and NPV by 9%. In regards to worsening of ambiguity and rule-in performance, the most affected score is FasterRisk, with an increase of 68% in ambiguity and a decrease of 49% in specificity. The least affected score is HEART, with an increase of 13% in ambiguity and a decrease of 25% in specificity. For the case of FasterRisk, we speculate the worsening of specificity is because the new cutoffs guarantee $FPR < 10\%$ which happens to be worse than the original FPR of 5%. In an attempt to rank all scores, we argue HEART+CC had the best discriminatory performance, followed by FasterRisk+CC and then by HEAR+CC. Even though HEAR+CC had the best rule-out performance, its coverage compromises practical applicability.

5.3. Results on Tabular Benchmark

Results on Table 3 suggest that the incorporation of class-conditional conformal (i.e. FasterRisk+CC) improved AUROC performance across 11 out of 12 datasets at the expense of coverage. FasterRisk’s coverage is 100% because it only employs one cutoff, as opposed to

1. <https://github.com/jjgarciac/cc-risk-stratification>

Method	Prevalence	Coverage	Sensitivity	Specificity	PPV	NPV	AUROC	Accuracy	AUPRC
FasterRisk	22	100	29	95	61	83	72	80	44
FasterRisk+CC	34	32	93	46	47	92	77	62	56
HEAR	20	85	20	89	30	82	65	75	27
HEAR+CC	19	19	93	50	30	97	72	58	30
HEART	15	54	42	89	40	90	74	82	36
HEART+CC	20	41	88	64	38	95	78	69	41

Table 2: MACE classification performance for each method (i.e. FasterRisk, HEAR, HEART) along with the use of class-conditional (+CC) conformal prediction to re-estimate cutoffs. In all cases we observe an increase in sensitivity with reduction in coverage.

CC, which employs two, and introduce an ambiguous region. Curiously, the addition of CC was detrimental to the Breast Cancer prediction. We conjecture this is because the constraints on FPR and FNR are larger than the FPR and FNR of the original cutoff.

6. Ablations

6.1. Are the theoretical guarantees met in other datasets?

Yes, we observe in Figure 3, the condition ($FPR \leq \alpha_- = 0.1$) is satisfied 96% of the time; similarly, the condition ($FNR \leq \alpha_+ = 0.1$) materializes 77% of the time. Lastly, we see the bound on error (Section 6.2) materializes 92% of the time. We speculate the variability observed is correlated with sample size and predictor performance. Similar variability is observed with other conformal estimators in Angelopoulos et al. (2023).

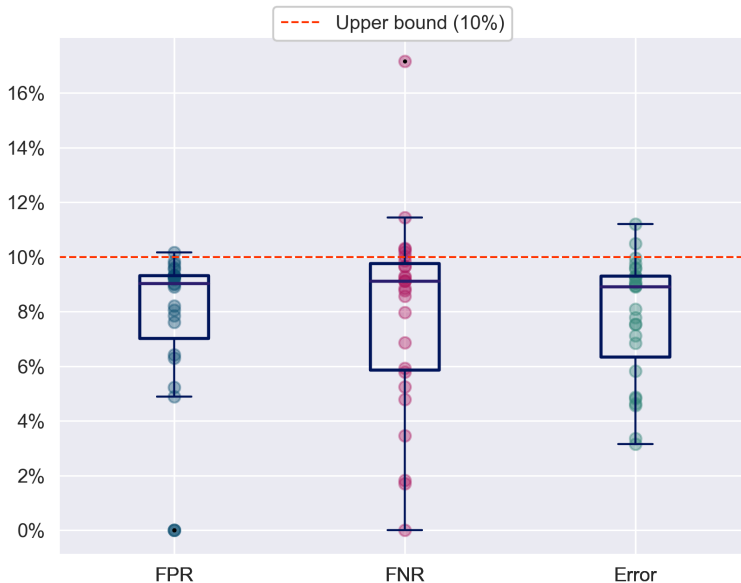


Figure 3: Boxplots for the observed FPR, FNR and Error as defined in Sections 6.2 and 6.1. Each point corresponds to a (fold, dataset) from Grinsztajn et al. (2022) listed in Table 3. Cutoffs were estimated to theoretically bound FPR, FNR and Error by 10%.

6.2. How does prevalence affect the choice of tolerance for FPR and FNR?

In terms of error (i.e. incorrectly labeling positive cases as low-risk or negative cases as high-risk), it is better to have more stringent control over the most prevalent class (e.g. If the most prevalent class is $-$ then smaller α_- mitigates more errors). This follows from:

$$P(\text{Error}) = \text{FNR} \cdot P(Y = +) + \text{FPR} \cdot P(Y = -) \leq \alpha_+ P(Y = +) + \alpha_- P(Y = -)$$

However, if there is a preference for identification of the least prevalent class it is important to consider the tradeoff between ambiguity and control. For instance, in the case of MACE rule-out performance, it is better to be more stringent on the FNR than the FPR. Nonetheless, due to low prevalence of MACE, this control strategy results in considerably more false positives than false negatives.

Dataset	Method	Coverage (%)	AUROC (%)	Prevalence (%)
compas-two-years (n=4966, d=11)	FasterRisk	100 ± 0.0	73 ± 0.7	50 ± 0.0
	FasterRisk+CC	40 ± 0.4	80 ± 0.5	53 ± 1.3
california (n=20634, d=8)	FasterRisk	100 ± 0.0	86 ± 3.5	50 ± 0.0
	FasterRisk+CC	57 ± 6.2	92 ± 3.5	54 ± 5.2
albert (n=58252, d=31)	FasterRisk	100 ± 0.0	68 ± 0.5	50 ± 0.0
	FasterRisk+CC	35 ± 0.9	74 ± 0.2	48 ± 0.4
bank-marketing (n=10578, d=7)	FasterRisk	100 ± 0.0	60 ± 13.4	50 ± 0.0
	FasterRisk+CC	20 ± 11.8	66 ± 13.1	48 ± 19.7
MagicTelescope (n=13376, d=10)	FasterRisk	100 ± 0.0	82 ± 0.7	50 ± 0.0
	FasterRisk+CC	57 ± 1.7	86 ± 0.9	40 ± 0.1
house.16H (n=13488, d=16)	FasterRisk	100 ± 0.0	68 ± 0.4	50 ± 0.0
	FasterRisk+CC	27 ± 0.6	61 ± 0.5	17 ± 0.1
heloc (n=10000, d=22)	FasterRisk	100 ± 0.0	75 ± 1.6	50 ± 0.0
	FasterRisk+CC	45 ± 0.9	81 ± 1.5	52 ± 1.1
default-of-credit...(cat.) (n=13272, d=21)	FasterRisk	100 ± 0.0	70 ± 1.5	50 ± 0.0
	FasterRisk+CC	36 ± 2.0	77 ± 0.6	71 ± 1.7
default-of-credit...(num.) (n=13272, d=20)	FasterRisk	100 ± 0.0	70 ± 1.4	50 ± 0.0
	FasterRisk+CC	36 ± 1.1	78 ± 0.7	71 ± 2.3
Mammo (n=961, d=14)	FasterRisk	100 ± 0.0	85 ± 6.4	46 ± 0.2
	FasterRisk+CC	68 ± 9.1	87 ± 7.4	37 ± 4.3
Spambase (n=4601, d=57)	FasterRisk	100 ± 0.0	90 ± 1.9	39 ± 0.1
	FasterRisk+CC	67 ± 9.7	94 ± 0.6	40 ± 2.5
BreastCancer (n=683, d=9)	FasterRisk	100 ± 0.0	99 ± 0.6	35 ± 0.5
	FasterRisk+CC	88 ± 7.0	100 ± 0.4	34 ± 3.3

Table 3: Performance evaluation of FasterRisk with and without conformal prediction (CC) across twelve datasets. The addition of CC notably enhances AUROC scores, indicating improved discriminatory power, albeit with a trade-off in coverage.

6.3. Could HEAR/HEART satisfy the most acceptable performance according to cardiologists?

No. According to [Cooper et al. \(2023\)](#), clinicians report that 99% sensitivity offers safe rule-out, a goal not yet attained by a risk score. To attempt this goal, we set $\alpha_+ = 0.01$

and theoretically bound $FNR \leq 1\%$. Unfortunately, the estimated cutoff q_+ , given this constraint, yields only scores of 0 as low risk and thus loses practical utility (see Figure 4). This analysis suggests that neither the HEAR, nor HEART scores are sufficient to sensibly satisfy the 99% sensitivity goal in our sample population and thus other scores are needed (e.g. FasterRisk, TIMI, ECADS). In general, class-conditional conformal estimation provides evidence, through the cutoffs, that a predictor may be insufficiently discriminative for a particular population. That is, a large portion of cases are considered ambiguous or intermediate risk.

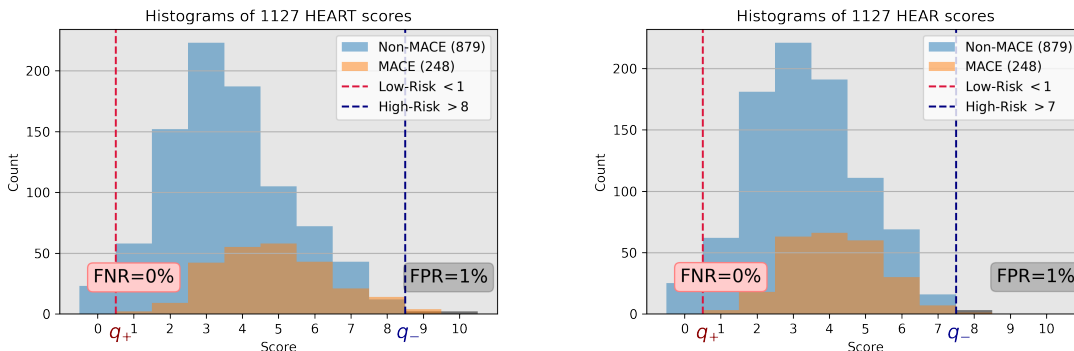


Figure 4: Histogram of test scores per class (Setup similar to Figure 1) for HEART (Left) and HEAR (Right). The blue and red lines correspond to cutoffs that bound the $FNR \leq 1\%$ and $FPR \leq 1\%$ respectively. For both scores, the data population suggest $q_+ = 1$ to satisfy the sensitivity guarantee imposed in Cooper et al. (2023). Accordingly, we conclude neither score is sufficient and better discriminators are required to further mitigate errors.

7. Discussion

This study is the first to evaluate the utility of class-conditional conformal estimation as a mechanism to enhance the stratification performance of risk scores. Our empirical results on binary classification of several (datasets, method) pairs evidenced an AUROC improvement in 14 out of 15 cases. Furthermore, we observed that the theoretical guarantees materialize 96% of the time for the FPR and 77% of the time for the FNR. These results reassure both practitioners and clinicians that FPR and FNR limits (i.e. avoiding over-treatment or missing cases) will not be exceeded and that the newly estimated cutoffs will mitigate stratification errors.

In the task of 30-day MACE prediction, we observed an improvement in the rule-out performance of the HEAR score and standard of care HEART score. Concretely, the HEART score sensitivity and NPV increased, by 46% and 5% respectively, with a compromise of 13% in ambiguity and 25% in specificity. The HEAR score sensitivity and NPV increased, by 73% and 15% respectively, with a compromise of 13% in ambiguity and 39% in specificity. Given that missed diagnosis may result in irreversible damage to the myocardium, it is reasonable to increase the number of intermediate-risk cases that require further evaluation. Nonetheless, a considerable increase in the proportion of ambiguous cases signals the score

has limited utility in a given population. The challenge lies in controlling the FPR and FNR without compromising the model’s coverage beyond a sensible limit. To choose the FPR and FNR, it is beneficial in terms of accuracy to be less tolerant of errors in the most prevalent class (e.g. smaller α_- for MACE prediction) noting this also increases ambiguity the most.

In conclusion, risk scores represent inherently explainable models used in real world decision making pipelines (e.g. Apgar for newborn evaluation, Grace for assessment of probability of death for discharge ACS patients, etc). Strategies like CC can mitigate stratification errors of such scores without compromising their explainability by tailoring the cutoffs to the characteristics and constraints of the deployment site.

Limitations The conformal approach requires both the cutoff-estimation-data and the test-data to be exchangeable. Therefore, it is not applicable in situations with a possible population shift between estimation and deployment populations. The conformal algorithm also requires known sensible limits for the FPR and FNR and respecting them may increase the proportion of ambiguous cases beyond a useful limit. Further comparisons with other concurrent cutoff estimation approaches like [Angelopoulos et al. \(2024\)](#) could provide more insight as to the clinical benefits of controlling PPV and NPV instead.

References

- Salah Al-Zaiti, Lucas Besomi, Zeineb Bouzid, Ziad Faramand, Stephanie Frisch, Christian Martin-Gill, Richard Gregg, Samir Saba, Clifton Callaway, and Ervin Sejdić. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature communications*, 11(1):3966, 2020.
- Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.
- Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.
- Anastasios Nikolas Angelopoulos, Stuart R Pomerantz, Synho Do, Stephen Bates, Christopher P Bridge, Daniel C Elton, Michael H Lev, R Gilberto Gonzalez, Michael I Jordan, and Jitendra Malik. Conformal triage for medical imaging ai deployment. *medRxiv*, pages 2024-02, 2024.
- BE Backus, AJ Six, JC Kelder, MAR Bosschaert, EG Mast, A Mosterd, RF Veldkamp, AJ Wardeh, R Tio, R Braam, et al. A prospective validation of the heart score for chest pain patients at the emergency department. *International journal of cardiology*, 168(3): 2153–2158, 2013.
- Vineeth Nallure Balasubramanian, R Gouripeddi, Sethuraman Panchanathan, J Vermillion, A Bhaskaran, and RM Siegel. Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. In *2009 36th Annual computers in cardiology conference (CinC)*, pages 5–8. IEEE, 2009.

- Jamie G Cooper, James Ferguson, Lorna A Donaldson, Kim MM Black, Kate J Livock, Judith L Horrill, Elaine M Davidson, Neil W Scott, Amanda J Lee, Takeshi Fujisawa, et al. Performance of a prehospital heart score in patients with possible myocardial infarction: a prospective evaluation. *Emergency Medicine Journal*, 40(7):474–481, 2023.
- Luigi Pietro Cordella, Claudio De Stefano, Francesco Tortorella, and Mario Vento. A method for improving classification reliability of multilayer perceptrons. *IEEE Transactions on Neural Networks*, 6(5):1140–1147, 1995.
- Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.
- Yonatan Elul, Aviv A Rosenberg, Assaf Schuster, Alex M Bronstein, and Yael Yaniv. Meeting the unmet needs of clinicians from ai systems showcased for cardiology with deep-learning-based ecg analysis. *Proceedings of the National Academy of Sciences*, 118(24):e2020620118, 2021.
- Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472, 2005.
- Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.
- Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.
- Ulf Johansson, Cecilia Sonstrod, Tuwe Lofstrom, and Henrik Bostrom. Confidence classifiers with guaranteed accuracy or precision. In *Conformal and Probabilistic Prediction with Applications*, pages 513–533. PMLR, 2023.
- Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What’s a good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.
- Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.
- Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer, and Cynthia Rudin. Fasterrisk: Fast and accurate interpretable risk scores. *Advances in Neural Information Processing Systems*, 35:17760–17773, 2022.
- Harris Papadopoulos, Efthymoulos Kyriacou, and Andrew Nicolaides. Unbiased confidence measures for stroke risk estimation based on ultrasound carotid image analysis. *Neural Computing and Applications*, 28:1209–1223, 2017.
- Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2):4, 2020.

- Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019.
- Jason P Stopyra, William S Harper, Tyson J Higgins, Julia V Prokesova, James E Winslow, Robert D Nelson, Roy L Alson, Christopher A Davis, Gregory B Russell, Chadwick D Miller, et al. Prehospital modified heart score predictive of 30-day adverse cardiac events. *Prehospital and disaster medicine*, 33(1):58–62, 2018.
- Masahiko Takeda, Takehiko Oami, Yosuke Hayashi, Tadanaga Shimada, Noriyuki Hattori, Kazuya Tateishi, Rie E Miura, Yasuo Yamao, Ryuzo Abe, Yoshio Kobayashi, et al. Prehospital diagnostic algorithm for acute coronary syndrome using machine learning: a prospective observational study. *Scientific Reports*, 12(1):14593, 2022.
- Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.
- Francesco Tortorella. An optimal reject rule for binary classifiers. In *Advances in Pattern Recognition: Joint IAPR International Workshops SSPR 2000 and SPR 2000 Alicante, Spain, August 30–September 1, 2000 Proceedings*, pages 611–620. Springer, 2000.
- Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.
- Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.
- Jessica K Zègre-Hemsey, Melanie Hogg, Jamie Crandell, Michele M Pelter, Len Gettes, Eugene H Chung, David Pearson, Pilar Tochiki, Jonathan R Studnek, and Wayne Rosamond. Prehospital ecg with st-depression and t-wave inversion are associated with new onset heart failure in individuals transported by ambulance for suspected acute coronary syndrome. *Journal of electrocardiology*, 69:23–28, 2021.