Statistica Sinica Preprint No: SS-2025-0127

Title	Classification Uncertainty Quantification: A Comparison Between Bootstrap and Conformal ROC Confidence Bands
Manuscript ID	SS-2025-0127
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0127
Complete List of Authors	Zheshi Zheng, Bo Yang and Peter Song
Corresponding Authors	Peter Song
E-mails	pxsong@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

Statistica Sinica

Classification uncertainty quantification: A comparison between bootstrap and conformal ROC confidence bands

Zheshi Zheng, Bo Yang, Peter X.-K. Song

Department of Biostatistics, University of Michigan

Abstract: Receiver Operating Characteristic (ROC) curves are commonly used to evaluate the performance of classification or prediction algorithms. However, in the literature, the uncertainty assessment of such algorithm is less rigorously addressed. In this article, we examine the limitations of the popular bootstrap method for ROC uncertainty quantification, elucidated by a simple yet essential model-based classification approach. We show that our proposed approach based on conformal prediction provides a valid solution for quantifying the uncertainty of the ROC curve and the Youden index. Both theoretical and numerical results corroborate the improved uncertainty quantification by the conformal inference over the bootstrap method.

Key words and phrases: Logistic model, model-based classification, prediction inference, Youden Index.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

1. Introduction

Classification as an important supervised learning approach is undertaken pervasively in various applied studies. To evaluate performance of a classification algorithm, the Receiver Operating Characteristic (ROC) curve (Fawcett, 2006) is the choice of the analytic for such an evaluation task. While there is vast literature concerning the ROC curve estimation, uncertainty quantification in connection to bias, variability and stability of a chosen classifier has been little investigated. From the point of view on statistical learning, evaluating the uncertainty of a classification algorithm as part of its performance is arguably of critical importance. In the literature, this evaluation task via the bootstrap method has been taken for granted with no or little theoretical justification. In this article, we revisit the popular bootstrap-based uncertainty quantification for ROC curves to unveil some key limitations that are largely ignored in the current practice. To overcome the limitations, we propose a conformal prediction based solution, which offers a valid and robust approach to quantifying the uncertainty of the ROC curve in comparison to the popular bootstrap method.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

1.1 Problem setting and ROC curve

We consider a classification problem with a dataset of independent observations $\mathcal{D}_{obs} = \{(y_i, z_i)\}_{i=1}^n$ from n subjects consisting of label variable $y_i \in \{0,1\}$ and a p'-element vector of numerical features, denoted by $z \in \mathbb{R}^{p'}$ and indexed by $J = \{1, 2, \cdots, p'\}$. Let $x = z_{J^*} \in \mathbb{R}^p$ denote the set of p important features with $J^* \subseteq J$. To evaluate the performance of a classification algorithm, we split the observed data into three sets $\mathcal{D}_{obs} = \mathcal{D}_{tr} \cup \mathcal{D}_{ca} \cup \mathcal{D}_{tst}$; the first dataset $\mathcal{D}_{tr}$ is used to train a classification algorithm; the second dataset $\mathcal{D}_{ca}$ is used for algorithm calibration, followed by the performance evaluation using the third dataset $\mathcal{D}_{tst}$ (*a.k.a.* the *test dataset*). We consider a simple yet popular model-based classifier via the logistic regression to first elucidate limitations of the widely used bootstrap method for uncertainty quantification. Then, in the same setting, we illustrate the validity of a conformal prediction based method proposed in this paper to overcome the limitations. A direct comparison of these two uncertainty quantification analytics are carried out on the test dataset $\mathcal{D}_{tst}$.

Suppose that there exists a true data generative model for observed labels:

$$y \sim Ber\big(\pi^*(x)\big) \text{ with } \eta^*(x) = logit\{\pi^*(x)\} = x^T\beta^*, \tag{1.1}$$

where $\pi^*(x) = P(y = 1|x)$. Clearly, this true label prediction model involves only the "oracle" x-features. In practice, the true model is unknown and needs to be determined by identifying most, if not all, of these x-features from the z-features through either the domain knowledge or statistical variable selection methods using the training data $\mathcal{D}_{tr}$.

Denote the set of identified features as $w = z_{J_{tr}} \in \mathbb{R}^q$ with $J_{tr} \subseteq J$, obtained at the model training stage on data $\mathcal{D}_{tr}$, which gives rise to a classifier with the coefficient β_{tr}. The resulting classification algorithm then outputs a prediction value of the linear predictor $\eta_{tr}(w_{new}) = w_{new}^T \beta_{tr}$ for any new input feature w_{new}. This consequently produces a predicted probability, π_{new} via the monotonic expit (i.e. the inverse of logit) function transformation. Moreover, the performance of this probability prediction algorithm is routinely assessed with the utility of ROC curves.

Our objective focuses on uncertainty quantification as a critical part of the ROC-based performance evaluation specific to a given classifier η_{tr} through the means of confidence bands. Visualizing an ROC curve of classifier $\eta_{tr}(\cdot)$ is done by plotting the sensitivity against one minus specificity defined as follows, respectively; that is, for a cutoff $c \in \mathbb{R}$,

$$\text{Sens}(c) = P\{\eta_{tr}(w) > c \mid y = 1\}; \quad \text{Spec}(c) = P\{\eta_{tr}(w) \le c \mid y = 0\}. \quad (1.2)$$

In practice, we calculate their sample counterparts in the estimation of

ROC curve, namely the empirical sensitivity and specificity of the following forms:

$$\hat{Sens}(c) = \sum_{j \in \mathcal{I}_{tst,1}} \frac{1\{\eta_{tr}(w_j) > c\}}{n_{tst,1}}; \quad \hat{Spec}(c) = \sum_{j \in \mathcal{I}_{tst,0}} \frac{1\{\eta_{tr}(w_j) \le c\}}{n_{tst,0}}, \tag{1.3}$$

where $\mathcal{I}_{tst,k} = \{j : y_j = k, \forall j \in \mathcal{D}_{tst}\}$ with $\mathcal{I}_{tst,1} \cup \mathcal{I}_{tst,0} = \mathcal{D}_{tst}$ and $n_{tst,k} = |\mathcal{I}_{tst,k}|, k = 0, 1$. Throughout the paper, $1(\cdot)$ denotes the indicator function. A plot of empirical ROC curve is made by interpolating pairs of $\left(1 - \hat{Spec}(c), \hat{Sens}(c)\right)$ over grid values $c \in \mathbb{R}$.

Adding confidence bands of an empirical ROC curve has been considered to reflect the variability of classification algorithm's performance. Jensen et al. (2000) provides a review of several methods to construct the confidence bands for ROC curves, including the point-wise confidence bands (Schafer, 1994; Hilgers, 1991) and global confidence bands (Campbell, 1994), among others. In particular, the utility of diagnostic testing (Nakas et al., 2023) to construct point-wise ROC confidence bands is predominant in recent developments for uncertainty quantification. However, the assumption that the distributions of $\eta_{tr}(w)$ given $y = 1$ and $y = 0$ are independent and take certain parametric forms, which is used for uncertainty quantification in the diagnostic testing literature, are not applicable to a general classification setting. More importantly, as shown in Seciton 1.2, the popular bootstrap based uncertainty quantification (Adler and Lausen, 2009) has some noticed drawbacks that may lead to inappropriate uncertainty quantification.

1.2 Point-wise bootstrap ROC confidence bands

The point-wise confidence bands for a ROC curve are of practical importance as the uncertainty at a certain cutoff point is often of interest in assessing the performance of a classification algorithm. Technically, we need to construct confidence intervals for both $Sens(c)$ and $Spec(c)$ at a given threshold $c \in \mathbb{R}$. To gain insights on the performance of bootstrap method for such a task, below we focus on a linear classifier $\eta_{tr}(w) = w^T \beta_{tr}$ and constructing confidence bands for its sensitivity. A similar uncertainty analysis may be extended to specificity with little effort.

Here we adopt the threshold averaging (TA) method (Fawcett, 2006) that constructs a confidence interval for $Sens(c)$ at a fixed c. Refer to Algorithm 1 for the detailed steps of bootstrap-t algorithm introduced by DiCiccio and Efron (1996). This type of bootstrap method is commonly used in practice with a well-developed R package `pROC`. We also consider other types of bootstrap methods such as the so-called double bootstrap through resampling on both training and test datasets. We report the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

detailed comparisons of these bootstrap methods with our proposed method in the simulation study in the Supplementary S2.1.

Following the classical bootstrap theory, we can show under mild regularity conditions: With a large bootstrap sample size B, the difference between two distributions of sensitivity satisfies $\sup_t |\hat{F}_B(t) - F_n(t)| = o_p\big(n_{tst,1}^{-1/2}\big)$ where $\hat{F}_B(t) = \frac{1}{B}\sum_{b=1}^{B} 1\big[\sqrt{n_{tst,1}}\{\hat{Sens}^{b}(c) - \hat{Sens}(c)\} \le t\big]$, $F_n(t) = P\big[\sqrt{n_{tst,1}}\{\hat{Sens}(c) - Sens(c)\} \le t\big]$. Consequently, the asymptotic coverage is given by

$$P\big\{Sens(c) \in CI_B(c;\alpha)\big\} = 1 - \alpha - o(n_{tst,1}^{-1/2}),$$

where $CI_B(c;\alpha)$ provides a suitable $100(1-\alpha)\%$ bootstrap confidence interval under the bootstrap empirical distribution $F_B(t)$. It is worth noting that this kind of bootstrap confidence interval only quantifies the variance of $\hat{Sens}(c)$ but ignores potential $bias = \eta_{tr}(w) - \eta^*(x)$ occurred in the training stage. A proper uncertainty quantification is expected to account for both bias and variance of a given classifier $\eta_{tr}(w)$. To discuss further, we first define the target estimands, also termed as "oracle" sensitivity and specificity in this paper, as follows:

$$Sens_0(c) = \textstyle\sum_{j\in\mathcal{I}_{tst,1}} \frac{1\{\eta^*(x_j)>c\}}{n_{tst,1}}; \quad Spec_0(c) = \textstyle\sum_{j\in\mathcal{I}_{tst,0}} \frac{1\{\eta^*(x_j)\le c\}}{n_{tst,0}}. \tag{1.4}$$

To elucidate the need of these oracle estimands in the uncertainty quantification, we design a simple simulation study to observe the behavior of the bootstrap confidence bands from three classifiers that are subject to different degrees of model mis-specification. We generate three data sets of random samples, $\mathcal{D}_{tr}, \mathcal{D}_{ca}$ and $\mathcal{D}_{tst}$, each of size 500, from a logistic data generating model (GM) $\eta^*(x) = x_1+1.4x_2+1.8x_3$ where the three covariates are independently drawn from the standard normal $N(0, 1)$. We construct the bootstrap confidence bands using Algorithm 1 for three classifiers: (TM1) $\eta_{tr,1} = x_1\beta_{1,tr} + x_2\beta_{2,tr} + x_3\beta_{3,tr}$, (TM2) $\eta_{tr,2} = x_2\gamma_{2,tr} + x_3\gamma_{3,tr}$, and (TM3) $\eta_{tr,3} = x_1\xi_{1,tr} + x_2\xi_{3,tr}$. Of note, models TM2 and TM3 are different from the true data generation model GM, and TM3 ignores the strongest feature while TM2 misses the weakest feature. Following the standard model training procedure, we estimate the respective parameter vectors $\beta_{tr}, \gamma_{tr}, \xi_{tr}$ using the data $\mathcal{D}_{tr} \cup \mathcal{D}_{ca}$ as the step of calibration is not needed here.

The resulting ROC curves and their bootstrap confidence bands (from 1000 bootstrap samples) for the three classifiers are plotted, respectively, in the left column of Figure 1. Clearly, with no model mis-specification, TM1-classifier produces an ROC curve, termed as TM1-specific ROC curve, that closely matches the oracle curve constructed under the true data-generating model with the true parameter values. Its bootstrap ROC confidence bands cover both the TM1-specific ROC curve and the oracle ROC curve, as

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

expected. The performance of the TM2-classifier is worse, with the TM2-specific ROC curve slightly deviating from the oracle one, yet the bootstrap ROC confidence bands only cover the TM2-specific ROC curve and fail to cover the oracle curve. For the TM3 classifier, its performance is the worst, with the TM3-specific ROC curve being significantly distance from the oracle curve, and the bootstrap confidence band not covering the oracle ROC. An important insight from this simulation study is that all three bootstrap confidence bands appear remarkably narrow, regardless of the degree of model mis-specification. Such lack of responsiveness to classifier quality raises concerns with appropriateness and usefulness of bootstrap confidence bands in practical studies in that most of time classifier are based on mis-specified models. In contrast, as shown in the right column of Figure 1, the conformal ROC confidence bands developed in this paper effectively respond to classifiers of varying performance quality. Due to this demonstrated responsiveness, conformal ROC confidence bands provide a more desirable quantification of data uncertainty for a good or bad classification rule compared to the bootstrap method.

Conceptually, poor classifiers are expected to receive wide confidence bands in order to ensure a proper coverage rate due to the disturbance of estimation bias, whereas narrow bands are expected for stronger classifiers.

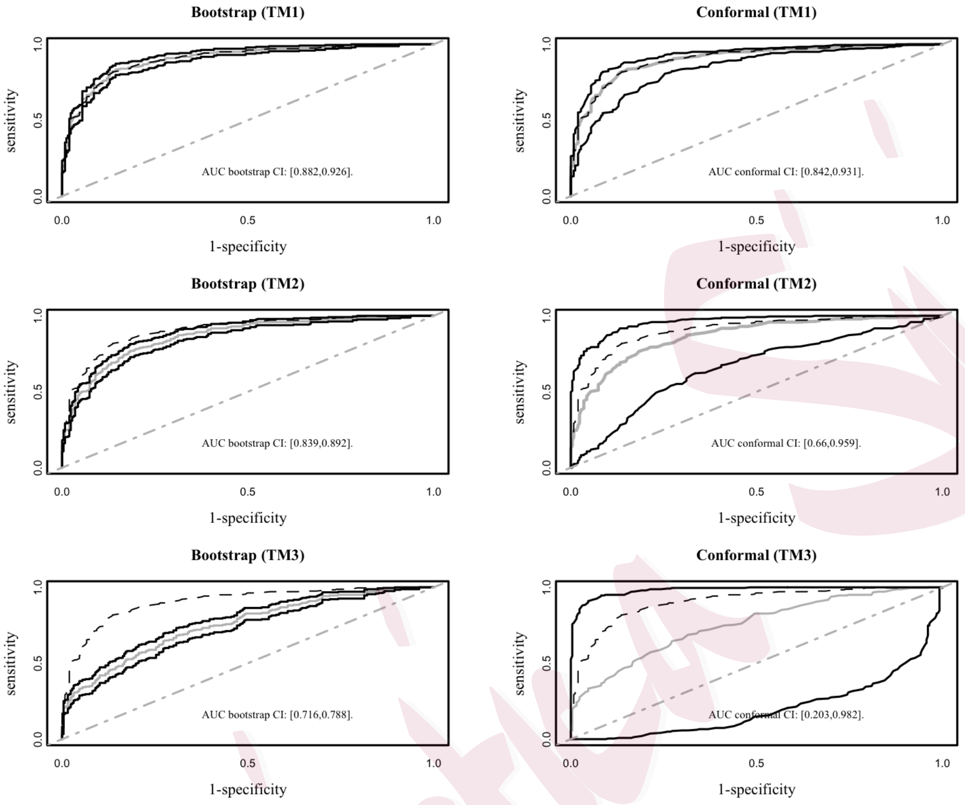

Figure 1: ROC confidence bands from bootstrap and conformal methods for (TM1) - (TM3) classification models. The solid gray line is the ROC curve from the classification model; the dashed gray line is the oracle ROC curve from the oracle data generating model; the two solid black lines are the confidence bands of the ROC curves; and the dot-dashed gray line is the reference line of $y = x$ (i.e. the ROC curve of random guess). The AUCs of the three classifiers are $0.907, 0.877, 0.743$ respectively, and the oracle AUC is 0.907.

In Example 1 below, analytically, we demonstrate a contradiction in the bootstrap confidence bands: a correctly specified classification model may produce wider bootstrap confidence intervals than a mis-specified model at some points on the ROC curve. This raises a concern about suitability and reliability of the bootstrap method in practical applications.

Example 1. Consider a classification model (1.1) with independent p-dimensional important features $x = (x_1, \dots, x_p)^T \sim N(0, I_p)$. Let $J^* = \{1, 2, \cdots, p\}$, $p \ll n$ and $x_{-1} = x_{J^* \setminus \{1\}}$. Suppose that in the construction of the first classifier, denoted by [TM$_\dagger$], through providence we perfectly selected the true features x_{J^*} constituting the set of identified features w, while in the construction of the second classifier denoted by [TM$_\ddagger$], we failed to identify one important feature, *say* x_1. We write these two models as [TM$_\dagger$]: $\eta_{tr,1}(w) = x^T \beta_{tr}$, and [TM$_\ddagger$]: $\eta_{tr,2}(w) = x_{-1}^T \gamma_{tr}$. Clearly, [TM$_\ddagger$] presents under-fitting. At a given threshold c, the length of a $100(1-\alpha)\%$ bootstrap confidence interval on the training data $\mathcal{D}_{tr}$, denoted by $CI_B(c;\alpha)$, is given by

$$\begin{aligned}\mathrm{LEN}\{CI_B(c;\alpha)\} &= n_{tst,1}^{-1/2}\left\{\hat{F}_B^{-1}(1-\frac{\alpha}{2}) - \hat{F}_B^{-1}(\frac{\alpha}{2})\right\}\\ &= n_{tst,1}^{-1/2}\left\{F_n^{-1}(1-\frac{\alpha}{2}) - F_n^{-1}(\frac{\alpha}{2})\right\} + o_p(n_{tst,1}^{-1})\\ &= z_\alpha Var^{1/2}\Big[1\{\eta_{tr}(w) > c \mid y = 1\}\Big] + o_p(n_{tst,1}^{-1}) \qquad (1.5)\end{aligned}$$

where F_B and F_n are the empirical distribution of $\sqrt{n_{tst,1}}\big\{\hat{Sens}^b(c) - \hat{Sens}(c)\big\}$ and the distribution of $\sqrt{n_{tst,1}}\big\{\hat{Sens}(c) - Sens(c)\big\}$, respectively. Equality(1.5) holds by the fact that given $\mathcal{D}_{tr}$, $n_{tst}^{-1/2}\big\{\hat{Sens}(c) - Sens(c)\big\}$ converges weakly to a normal distribution, and z_α denotes the corresponding $(1-\alpha)$ normal quantile. Proposition 1 below compares the lengths of bootstrap confidence intervals obtained, respectively, by the two logistic model-based classifiers above.

Proposition 1. *Consider two classifiers, $\eta_{tr,1}(w)$ and $\eta_{tr,2}(w)$, defined in Example 1. There exists an interval $C' \subset \mathbb{R}$ such that for any $c \in C'$ the odd ratio for the two classifiers defined above satisfies*

$$OR(c) = \frac{P\big\{\eta_{tr,1}(w) > c, y = 1\big\}P\big\{\eta_{tr,1}(w) \leq c, y = 1\big\}}{P\big\{\eta_{tr,2}(w) > c, y = 1\big\}P\big\{\eta_{tr,2}(w) \leq c, y = 1\big\}} > 1. \tag{1.6}$$

Equations (1.6) and (1.5) are connected by the fact that the variance in (1.5) is proportional to $P\big\{\eta_{tr,k}(w) > c, y = 1\big\}P\big\{\eta_{tr,k}(w) \leq c, y = 1\big\}, k = 1, 2$. This proposition, even though obtained in a hypothetical case, shows that the bootstrap confidence interval from the optimal classifier [TM$_\dagger$] can be wider than that from the under-fitting classifier [TM$_\ddagger$] at some positions of the ROC curves. This result is indeed counterintuitive and presents contradicting evidence to the validity of the bootstrap method in establishing proper uncertainty quantification on the ROC curve. This motivates us to

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

explore a conformal method as a remedy to such a drawback.

2. Conformal ROC confidence band

2.1 Conformal ROC confidence band

Conformal prediction (Vovk et al., 2005) is a robust prediction inference method that constructs prediction intervals for individual data points in a test set with valid coverage, even if the working prediction model has undesirable prediction performance. In the recent literature, split conformal and Jackknife-plus conformal procedures have gained their popular (Lei et al., 2018). Despite its theoretically guaranteed coverage, a better trade-off between prediction bias and prediction uncertainty is essential. This is because a poor working prediction model often produces wide confidence intervals (cf. Xie and Zheng, 2022), and often they appear too wide to be useful in practical studies. In the context of classification analysis, the pursuit of a good classifier with small errors is attractive, as it produces narrow prediction intervals with a low amount of uncertainty.

Our goal is to apply the conformal prediction to construct confidence intervals for sensitivity and specificity defined in (1.4). Of note, the batch conformal inference introduced in Lee et al. (2024) may be a method of choice for this task. However, based on our simulation results (see the last part of Table S2 in the Supplementary Material) this method appears to be computationally costly. To facilitate scalability and improve computational efficiency, we take a different way to construct confidence intervals, as detailed below. We first establish individualized prediction intervals for the classifier $\eta_{tr}(w)$ for every characteristic vector w in the test data $\mathcal{D}_{tst}$, and then combine these intervals according to their empirical distributions. The resulting conformal intervals reflect the overall quality of a classifier, in the hope that a better classifier produces narrower confidence bands. Thus, it seems natural to use the width of the ROC confidence band as a criterion to characterize the performance of a classifier.

Consider the conformity scores $R_i = \eta^*(x_i) - \eta_{tr}(w_i)$ for $i \in \mathcal{I}_{ca}$ (the i-th calibration data case) under the true classifier $\eta^*(x)$ and a built classifier $\eta_{tr}(w)$ from the training data. Obviously, $\eta^* = \eta^*(x), x \in \mathcal{D}_{tst}$ is an unknown "ideal", which is replaced in practice by a certain approximate "proxy", *say*, η. Given the notation of the conformity score $R(\eta) = \eta - \eta_{tr}(w)$, we obtain a set of ideal conformal scores $\{R_i, i \in \mathcal{I}_{ca}\}$ under $\eta = \eta^*(x)$ and another set of proxy conformity scores $\{\hat{R}_i = R_i(\eta), i \in \mathcal{I}_{tst}\}$ under $\eta = \hat{\eta}^*(w)$. Under the ideal situation with $\eta = \eta^*(x)$, with *i.i.d.* data, the ideal conformity scores $\{R_i, i \in \mathcal{I}_{ca}\}$ are exchangeable. It follows that an empirical p-value of a proxy η in the test data is given by

$$p_{ca}(\eta) = p(\eta|\mathcal{D}_{ca}) = \tfrac{1+\sum_{i\in\mathcal{I}_{ca}} 1\{R(\eta)\geq R_i\}}{|\mathcal{D}_{ca}|+1}, \text{ for some } \eta \text{ trained on } \mathcal{D}_{tst}, \quad (2.1)$$

and a $100(1-\alpha)\%$ prediction interval for this proxy η takes the form:

$$\{w : 2\min\{p_{ca}(\eta), 1-p_{ca}(\eta)\} \geq \alpha\} = \left[\eta_{tr}(w) + q^{ca}_{\alpha/2}(R_i), \eta_{tr}(w) + q^{ca}_{1-\alpha/2}(R_i)\right], \quad (2.2)$$

where the lower $q^{ca}_{\alpha/2}(R_i)$ and upper $q^{ca}_{1-\alpha/2}(R_i)$ are, respectively, the $\big[\alpha(|\mathcal{D}_{ca}|+1)/2-1\big]$-th and $[(1-\alpha/2)(|\mathcal{D}_{ca}|+1)]$-th order statistics for the set of the ideal scores $\{R_i\}_{i\in\mathcal{I}_{ca}}$. Here $[a]$ denotes the largest integer that does not exceed a. In practice, replacing R_i in (2.2) by $\hat{R}_i = \hat{\eta}^*(w_i) - \eta_{tr}(w_i), i \in \mathcal{I}_{ca}$ leads to a predition interval of the following form:

$$PI(w,\alpha) = \left[\eta_{tr}(w) + q^{ca}_{\alpha/2}(\hat{R}_i), \eta_{tr}(w) + q^{ca}_{1-\alpha/2}(\hat{R}_i)\right], \quad (2.3)$$

where the lower and upper quantiles are similarly calculated from the proxy scores $\{\hat{R}_i\}_{i\in\mathcal{I}_{ca}}$ as done in (2.2). In this paper, we establish the theoretical guarantee for the coverage of $PI(w,\alpha)$ in Lemma B1 in Appendix 2, under the condition that $\hat{\eta}^*(w)$ is a weakly consistent estimator of the true $\eta^*(x)$ with w and x being the features measured on the same sampling unit in the test data. Under such a theoretical guarantee, we construct the confidence intervals of sensitivity by combining individual prediction intervals in the sub-data with only class label 1, denoted by $\mathcal{D}_{tst,1} = \{(y_i, x_i, z_i) \in \mathcal{D}_{tst} :$

$y_i = 1\}$. Plugging the upper and lower bounds of individual prediction intervals $PI(w_j, \alpha), w_j \in \mathcal{D}_{tst,1}$ given by (2.3) into (1.3), we construct the conformal confidence interval for sensitivity $Sens(c)$ in (1.2) at a threshold c as follows:

$$CI_{cf}(c, \alpha) = \big[\tfrac{1}{|\mathcal{I}_{tst,1}|}\textstyle\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{lo}(\alpha) > c\}, \tfrac{1}{|\mathcal{I}_{tst,1}|}\textstyle\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{up}(\alpha) > c\}\big], \quad (2.4)$$

where $b_j^{lo}(\alpha) = \eta_{tr}(w_j) + q^{ca,1}_{\alpha/2}(\hat{R}_i)$ and $b_j^{up}(\alpha) = \eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i)$ are individual upper and lower limits. See Algorithm 2 for the implementation details. Moreover, the ROC confidence bands for the sensitivity are plotted of $CI_{cf}(c, \alpha)$ over value $c \in \mathbb{R}$ with gaps filled by interpolation.

2.2 Theoretical coverage guarantee

We now establish the theoretical coverage guarantee for the conformal ROC confidence bands given in (2.4). We assume that at least one of features in w is continuous. The following conditions are required to yield a weakly consistent classifier $\hat{\eta}^*$ derived with the training data $\mathcal{D}_{tr}$.

Condition 1. $\lim_{|\mathcal{D}_{tr}|\to\infty} \mathbb{E}[1\{|\hat{\eta}^*(w_i) - \eta^*(x_i)| > \delta_n\}] = 0$ for a certain sequence $\delta_n \to 0$.

Several popular classifiers satisfy Condition 1, including the logistic model-based classifier introduced in (1.1), k-nearest neighbor classifiers (Mack and Rosenblatt, 1979) and Neural-Network type classifiers (Liu et al., 2017).

Condition 2. Assume the ideal conformity scores $R_i = \eta^*(x_i) - \eta_{tr}(w_i)$ are continuous with $\mathbb{E}_{F_{R,1}}(|R_i|) < \infty$ for all $i \in \mathcal{I}_{ca,1}$, where $F_{R,1}(\cdot)$ is the cumulative distribution function (CDF) of the ideal conformity score $R_i, i \in \mathcal{I}_{ca,1}$.

Condition 2 is relatively mild by requiring the error of the built classifier $\eta_{tr}(\cdot)$ to have a finite absolute expectation. Under the above two regularity conditions, we establish the theoretical guarantee for the coverage of the proposed ROC confidence bands in (2.4). Let $o_m(1)$ denote a small term for an integer m such that $o_m(1) \to 0$ as $m \to \infty$.

Theorem 1. *Assume Conditions 1 and 2 hold. Let* $n_1 = \min\{|\mathcal{D}_{tr}|, |\mathcal{D}_{ca,1}|, |\mathcal{D}_{tst,1}|\}$. *Then, for* $\alpha \in (0, 0.5)$ *and any randomly chosen* $c \in \mathbb{R}$, *we have for a large* n_1,

$$P\left\{Sens_0(c) \in CI_{cf}(c, \alpha)\right\} \geq 1 - 2\alpha - o_{n_1}(1).$$

Theorem 1 implies that the conformal ROC confidence bands given in (2.4) have a proper point-wise coverage for the oracle ROC curve defined by (1.4). Theorem 2 below extends the above result to ensure that the ROC confidence bands $CI_{cf}(c, \alpha)$ in (2.4) can also properly cover the ROC curve $Sens(c)$ generated from a consistent classifier with a minor additional condition.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

Theorem 2. *Assume that Condition 1 and 2 hold and that the tails of the CDF* $F_{R,1}(\cdot)$ *in Condition 2 satisfy* $P\{F_{R,1}^{-1}(\alpha/2) > 0\} = o_{n_1}(1)$ *and* $P\{F_{R,1}^{-1}(1-\alpha/2) < 0\} = o_{n_1}(1)$. *Then, for any* $c \in \mathbb{R}$, *we have* $Sens(c) \in CI_{cf}(c,\alpha)$ *almost surely as* $n_1 \to \infty$.

Remark 1. The bootstrap ROC confidence bands guarantee a $100(1-2\alpha)\%$ coverage for $Sens(c)$ in (1.2), while the proposed conformal confidence bands $CI_{cf}(c,\alpha)$ in (2.4) guarantee an almost surely coverage as declared in Theorem 2. In addition, this same conformal ROC confidence bands in (2.4) also guarantees $100(1-\alpha)\%$ coverage for the oracle $Sens_0(c)$ in (1.4) according to Theorem 1. In contrast, the bootstrap confidence bands do not guarantee a proper coverage for the oracle $Sens_0(c)$. This insight makes an essential difference between the conformal method and the existing bootstrap method, and the latter typically yields a much narrower confidence bands than the former as shown in Figure 1.

2.3 Uncertainty quantification

With the available conformal ROC confidence bands $CI_{cf}(c,\alpha)$, we are ready to quantify the uncertainty of the ROC for a classification algorithm. Let us revisit the simulation in Figure 1, where such ROC confidence bands for training models TM1-TM3 are calculated from Algorithm 2, respectively, and displayed in the right column of the figure with the same simulated dataset. Although these bands cover both the model-based ROC curve and the oracle ROC, their performances are noticeably different in terms of their coverage ranges. The bands in model TM1 are the narrowest, with no surprise, as this prediction model is correctly specified, while the bands in model TM2 are slightly wider than the best, but remain above the diagonal line of random decision. The bands in model TM3 are the worst among the three cases and cover the diagonal line, meaning that there is no difference in the class label prediction between using model TM3 and flipping a fair coin. Such new ROC bands exhibit much higher sensitivity to the quality of classifier than the bootstrap confidence bands in the first column of Figure 1. Unfortunately, despite being widely used in practice, the bootstrap ROC confidence bands show a severe underestimation of uncertainty for mis-specified prediction models.

Extra simulation results comparing the coverage, length and the running time of the confidence bands from different bootstrap and conformal methods are provided in the Supplementary Material. These results corroborate our theoretical results, demonstrating that the proposed conformal ROC confidence bands ensure on-target coverage for the oracle sensitivity and specificity, while almost surely covering the sensitivity and speci-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

ficity of a working prediction model. In contrast, the bootstrap method can only achieve on-target coverage for the sensitivity and specificity when the prediction model is correctly specified; otherwise, it suffers undercoverage. Since the bootstrap method does not target the oracle sensitivity and specificity, the resulting confidence bands are deemed restrictive, short of interpretability and undesirable in uncertainty quantification. In addition, our proposed method also enjoys computational efficiency and scalability, making it appealing to handle large data sets in pratical studies.

2.4 Interval length and Youden index

A further utility of the conformal ROC confidence bands lies in the calculation of the interval length (LEN) at a given c. Moreover, it may be applied to calculate Youden's J index defined as the maximum distance of an ROC curve from the diagonal line (i.e. the random guess), which is obtained at the corresponding threshold c value (Youden, 1950; Nakas et al., 2023). The length of $CI_{cf}(c, \alpha)$ for sensitivity at a fixed c is given by

$$\begin{aligned}
\mathrm{LEN}\big\{CI_{cf}(c,\alpha)\big\} &= \frac{1}{|\mathcal{I}_{tst,1}|}\sum_{j\in\mathcal{I}_{tst,1}}\big[1\{b_j^{up}(\alpha)>c\}-1\{b_j^{lo}(\alpha)>c\}\big]\\
&= \frac{1}{|\mathcal{I}_{tst,1}|}\sum_{j\in\mathcal{I}_{tst,1}}1\big\{c-q^{ca,1}_{1-\alpha/2}(\hat{R}_i)\le \eta_{tr}(x_j)\le c-q^{ca,1}_{\alpha/2}(\hat{R}_i)\big\}\\
&= P\left\{\eta_{tr}(x_j)\in\big[c-q^{ca,1}_{1-\alpha/2}(\hat{R}_i), c-q^{ca,1}_{\alpha/2}(\hat{R}_i)\big]\mid y_j=1\right\}+o_{n_{tst,1}}(1),
\end{aligned}$$

where the last equality is due to the law of large number. This interval length may be used to reflect the responsiveness of the uncertainty quantification method in relation to the performance of a classifier, in hope that a better classifier should have a shorter length. We revisit in Example 1 and use [TM$_\dagger$] (i.e. the true prediction model), $\eta_{tr,1}(w)$, and a new model [TM$_{tr}$] to construct a new classifier $\eta_{tr,3}(w) = w^T\gamma_{tr}$ where the features $w = x_{J_{tr}}$ are selected from observed predictors x using the training dataset. Classifier $\eta_{tr,3}(w)$ is different from the $\eta_{tr,2}(w)$ under model [TM$_\ddagger$] with the first feature x_1 being omitted. Proposition 2 below shows that the length of the conformal ROC confidence band from [TM$_\dagger$] is almost surely shorter than, or at least equal to, that of [TM$_{tr}$], subject to stochastically ignorable margins of errors.

Proposition 2. *For two classification models [TM$_\dagger$] and [TM$_{tr}$] given above with all features $\{x_i, i \in \mathcal{I}_{tst}\} \overset{i.i.d.}{\sim} N(0,1)$, the lower and upper limits satisfy, for large n_1,respectively,*

$$q^{ca,1}_{\alpha/2}(\hat{R}_{i,1}) \geq q^{ca,1}_{\alpha/2}(\hat{R}_{i,2}) + o_{p,n_1}(1); \text{ and } q^{ca,1}_{1-\alpha/2}(\hat{R}_{i,1}) \leq q^{ca,1}_{1-\alpha/2}(\hat{R}_{i,2}) - o_{p,n_1}(1),$$

where $\hat{R}_{i,k} = \hat{\eta}^*(w_i) - \eta_{tr,k}(w_i), i \in \mathcal{I}_{tst,1}, k = 1([TM_\dagger]), 2([TM_{tr}])$ with $\hat{\eta}^*(w_i)$ being an consistent estimator of $\eta*$ satisfying Condition 1.

Here a random variable X of order $o_{p,n}$ means that $\lim_{n\to\infty} P(X >$

$\epsilon) = 0$ for any $\epsilon > 0$. A proof of Proposition 2 is given in Section 1.2 of the Supplementary Material.

The Youden index is a commonly used criterion, in addition to the area under the curve (AUC), as a summary of the ROC curve. It is defined as the maximum vertical distance between the ROC curve and the $y = x$ line. Specifically, the oracle Youden indices is defined as

$$J^* = max_c\{Sens_0(c) + Spec_0(c) - 1\}.$$

A similar approach to combining individual intervals may be employed here to yield confidence intervals for Youden index J. Specifically, at a $c^\dagger$ value corresponding to the maximum, treating the lower limit b_j^{lo} as a predictor for each $j \in \mathcal{I}_{tst,1}$, we obtain $\hat{Sens}(c^\dagger; b_1^{lo})$. For specificity, we treat the upper limit b_j^{up} as a predictor for each $j \in \mathcal{I}_{tst,0}$, leading to the lower (upper) bound for specificity(1$-$specificity), denoted as $\hat{Spec}(c^\dagger; b_0^{up})$. Similar calculations give upper bounds of sensitivity and specificity, denoted as $\hat{Sens}(c^\dagger; b_1^{up})$ and $\hat{Spec}(c^\dagger; b_0^{lo})$, respectively. Consequently, we propose a confidence interval for J^* as $CI_J(\alpha) = [\hat{Sens}(c^\dagger; b_1^{lo}) + \hat{Spec}(c^\dagger; b_0^{up}) - 1 - \alpha, \hat{Sens}(c^\dagger; b_1^{up}) + \hat{Spec}(c^\dagger; b_0^{lo}) - 1 + \alpha]$, in which additional length of 2α is added to guarantee the coverage for J^*. Proposition 3 below establishes the theoretical coverage of the proposed confidence interval $CI_J(\alpha)$ for the oracle Youden index J^*. The related technical details of the proof may be found in Section 1.3 of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

the Supplementary Materials.

Proposition 3. *Assume Conditions 1 and 2 hold. Then, as $n_1 \to \infty$ we have $J^* \in CI_J(\alpha)$ almost surely.*

3. Application in Prediction of Sexual Maturation Status

We demonstrate the conformal ROC confidence bands in an empirical study that aims to predict sexual maturation status among adolescents using data collected from the ELEMENT (Early Life Exposures in Mexico to Environmental Toxicants) cohort of children living in Mexico City (Perng et al., 2019). Existing findings in the literature have unveiled impacts of exposures to environmental toxic agents on the onset of children's precocious puberty (Euling et al., 2008; Emmanuel and Bokor, 2017). Precocious puberty in human growth and development is likely to cause adverse health outcomes in later life, including shorter adult height, emotional distress, increased risk of behavioral problems, and long-term health risks such as obesity and cancer (Carel and Léger, 2008). Tanner stage, *a.k.a.* Sexual Maturity Rate (SMR) (Marshall and Tanner, 1970), measures developmental phase of sexual maturation, and stage-4 or higher is deemed as full maturation, which is the status of interest ($y = 1$). In this analysis we predict girl's full maturation using DNA methylome (w) consisting of highdimensional beta values measured at 850K CpG sites. DNA methylation markers have been extensively used to characterize cellular aging process and to predict biological age (Horvath and Raj, 2018; McEwen et al., 2020).

This application uses a random sample of $n = 304$ girls with 134 girls with tanner stage ≤ 3 (*a.k.a.* partial maturation) and the rest 170 with full maturation. We focus on 2736 CpG sites from 94 genes preselected by the means of elastic-net penalized regression (McEwen et al., 2020) in the construct of existing PedBE biological clock for children. To evaluate the performance of the logistic prediction model for full maturation,we set up a test data of randomly selected 50 girls, on which the ROC curve is plotted. To mitigate the randomness in the choice of test data, we split the test data 20 times and report the average confidence intervals. The confidence bands for ROC curve using both the conformal and bootstrap methods are calculated and compared in terms of the area under curve (AUC). We report the averaged AUC confidence intervals using both methods in Table 1. For the conformal method, we only report the sensitivity confidence intervals and leave the specificity confidence intervals in Section 2.2 of the Supplementary Materials. We choose the 80% confidence in the following discussion because (i) this confidence level is known to be the minimal bar for statistical power analysis most frequently adopted in biomedical study

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

designs and (ii) clear patterns enhance the comparison. The rest plots of the ROC confidence bands with 90% and 95% confidence levels are given in Section 2.2 of the Supplementary Material. Results show that when using the conformal method, only the 80% confidence interval for AUC of sensitivity is above 0.5 (i.e. the AUC for a random flip), indicating that we have 80% confidence that the prediction model performs better than random guess for full sexual maturation. For partial maturation, however, conformal confidence intervals for AUC of specificity show the existence of large uncertainty in the prediction. Compared to the bootstrap confidence intervals, the confidence intervals for AUC of sensitivity are much narrower, as expected.

To visually show and compare the conformal and bootstrap confidence bands for the ROC curve, Figure 2 displays the averaged 80% ROC confidence bands for both sensitivity and specificity. This figure shows that the confidence bands for sensitivity do not contain the diagonal line, meaning that with 80% confidence the prediction for full sexual maturation is better than the decision by random guess. In contrast, the 80% specificity confidence bands contain the diagonal line, implying a large uncertainty in the prediction for partial maturation. In contrast, in both scenarios the bootstrap confidence bands do not cover the diagonal line.

$1-\alpha$	80%	90%	95%
Conformal CI	$[0.590, 0.917]$	$[0.377, 0.953]$	$[0.367, 0.964]$
Bootstrap CI	$[0.668, 0.835]$	$[0.642, 0.855]$	$[0.618, 0.871]$

Table 1: Average 80%, 90% and 95% confidence intervals (CI) for area under curve (AUC) using conformal method and bootstrap method.

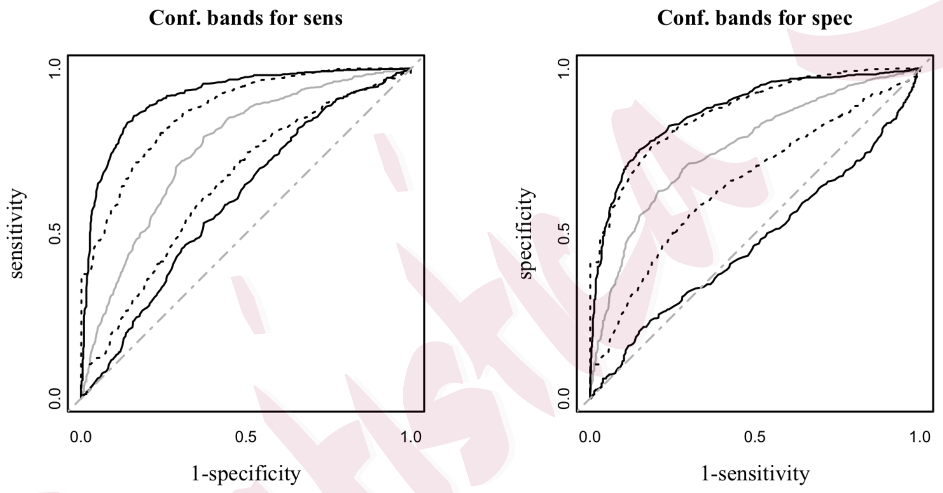

Figure 2: The 80% ROC confidence bands for sensitivity (left) and specificity (right) produced by conformal and bootstrap methods. Black lines are the ROC bands of full sexual maturation prediction; the solid lines are conformal confidence bands, while the dashed lines are bootstrap confidence bands. The gray solid line is the ROC curve of full sexual maturation prediction and the dot-dashed gray line is the reference line of random decision.

4. Discussion

Reliable uncertainty quantification is essential for evaluating classification models, as it offers a comprehensive understanding of how well a model generalizes to new data. In this article, we have studied uncertainty quantification methods for model-based classification algorithms, with a focus on the discussion on several limitations of popular bootstrap-based ROC confidence bands, including (i) the bands being too narrow so to underestimate uncertainty, (ii) the lack of variations in widths for classification models with varying quality of performance, and (iii) fail in covering the oracle ROC curve. To address these serious limitations, we proposed a new conformal inference based method that has been shown for its superiority in both theoretical and numerical results. We have illustrated that the proposed ROC confidence bands are particularly useful to provide reliable uncertainty quantification for the performance of mis-specified prediction models that are pervasive in practical studies.

There are several potential directions of future work to improve the conformal ROC confidence bands. First, the current solution is based on independent and identically distributed data, which may be violated in practice due, for example, to distributional shifts or correlations among sampling units. A future research direction of interest is to advance conformal prediction methods for conditional coverage, in which doubly robustness may be used to help relax the consistency requirement for the estimator $\hat{\eta}^*$. Second, confidence bands could be improved by improving the proxy classifier $\hat{\eta}^*(w)$, so improving the proxy conformal scores $\hat{R}_i$. The emerging technology of neural network modeling is promising to achieve this improvement. Thirdly, the theoretical results can be extended to accommodate a wider range of features such as functional, imaging, and text features. An immediate extension of interest is to generalize the method beyond binary classification to multi-label classification algorithms as Tanner's sexual maturation stage in our empirical study has 5 categories. Finally, we notice that our proposed construction of the confidence interval for Youden index J appears to be conservative, and an improvement with shorter interval lengths is worth a further exploration.

Supplementary Material

The Supplementary Material contains additional technical details concerning the proofs of Proposition 1 in Example 1, Proposition 2 and Proposition 3. Additional numerical results are also given in the Supplementary Material, including coverages for both bootstrap and conformal confidence bands at different confidence levels using simulated data, and supplemen-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

tary figures for the performances on the sexual maturation prediction. The R code used to generate the numerical results in this paper has been made publicly available at the GitHub repository:

https://github.com/ZheshiZheng/Conformal-ROC-confidence-band.

Acknowledgments

We thank Yahui Zhang for her help in the DNA methylation data used in the analysis. We thank the Associate Editor and two anonymous reviewers for their constructive comments and insightful suggestions, which have helped us significantly improve the paper. This research was supported by the U.S. National Institutes of Health R01ES033656.

Appendix

A Bootstrap and conformal algorithms for ROC confidence bands

Algorithm 1. Construction of a bootstrap confidence interval for $Sens(c)$ for a given $c \in \mathbb{R}$.

Step I: Train a linear classifier on training data and obtain $\eta_{tr}(w) = w^T \beta_{tr}$.

Step II: For $b \in \{1, 2, \cdots, B\}$, run the following loops.

Step II.a: Draw bootstrap sample $\mathcal{D}_{tst}^b = \{(y_j^b, x_j^b) : j \in \mathcal{I}_{tst}^b\}$ and set $\mathcal{I}_{tst,1}^b = \{j \in \mathcal{I}_{tst} : y_j^b = 1\}$.

Step II.b: Calculate $\hat{Sens}^b(c) = \frac{1}{n_{tst,1}} \sum_{j \in \mathcal{I}^b_{tst,1}} 1\{\eta_{tr}(w_j) > c\}$ where $n_{tst,1} = |\mathcal{I}^b_{tst,1}| = |\mathcal{I}_{tst,1}|$.

Step III: Compute the empirical CDF $\hat{F}_B(t) = \frac{1}{B} \sum_{b=1}^B 1\big[\sqrt{n_{tst,1}}\big\{\hat{Sens}^b(c) - \hat{Sens}(c)\big\} \leq t\big]$.

Step IV: Output $CI_B(c;\alpha) = \big[\hat{Sens}(c) - \hat{F}_B^{-1}(1-\alpha/2)/\sqrt{n_{tst,1}}, \hat{Sens}(c) - \hat{F}_B^{-1}(\alpha/2)/\sqrt{n_{tst,1}}\big]$.

Algorithm 2. Conformal confidence intervals for $Sens_0(c)$ for a given $c \in \mathbb{R}$ under a pre-specified proxy classifier $\hat{\eta}^*(w)$.

Step I: Train a classification algorithm on training data and obtain $\eta_{tr}(w) = w^T\beta_{tr}$.

Step II: For $j \in \mathcal{I}_{tst,1}$, run the following loops:

Step II.a: Calculate scores $\left\{\hat{R}_i = \hat{\eta}^*(w_i) - \eta_{tr}(w_i), i \in \mathcal{I}_{ca,1}\right\}$ where $\mathcal{I}_{ca,1} = \{j \in \mathcal{I}_{ca} : y_j = 1\}$;

Step II.b: Calculate $CI(x_j, \alpha) = [b_j^{lo}(\alpha), b_j^{up}(\alpha)]$,

where $b_j^{lo}(\alpha) = \eta_{tr}(w_j) + q^{ca,1}_{\alpha/2}(\hat{R}_i)$ and $b_j^{up}(\alpha) = \eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i)$.

Step III: Output confidence interval for $Sens_0(c)$ of the form:

$CI_{cf}(c,\alpha) = \big[\frac{1}{|\mathcal{I}_{tst,1}|}\sum_{j\in\mathcal{I}_{tst,1}} 1\big\{b_j^{lo}(\alpha) > c\big\}, \frac{1}{|\mathcal{I}_{tst,1}|}\sum_{j\in\mathcal{I}_{tst,1}} 1\big\{b_j^{up}(\alpha) > c\big\}\big]$.

B Proof of Theorem 1 and Theorem 2

We begin with some notations used in the proof. Denote $n_{ca} = |\mathcal{D}_{ca}|$, $n_{tst} = |\mathcal{D}_{tst}|$ and $n_{tst,1} = |\mathcal{D}_{tst,1}|$. Recall that $n_1 = \min\{n_{ca,1}, n_{tr}, n_{tst,1}\}$ and we define $n_2 = \min\{n_{ca}, n_{tr}\}$ and $n_3 = \min\{n_{ca,1}, n_{tr,1}\}$. For $j \in \mathcal{I}_{tst}$, denote $R_j = \eta^*(x_j) - \eta_{tr}(w_j) = x_j^T\beta^* - w_j^T\beta_{tr}$.

Lemma 1. *If Condition 1 holds, then for $\alpha \in (0, 0.5)$ and any $(y, x) \in \mathcal{D}_{tst}$,*

$$\lim_{n_2 \to \infty} P\left\{\eta^*(x) \in PI(w, \alpha)\right\} \geq 1 - \alpha,$$

where $PI(w, \alpha)$ is given in (2.3).

Proof of Lemma 1. Define the empirical CDFs $\hat{G}_{n_{ca}}(r) = \frac{1}{n_{ca}} \sum_{i \in \mathcal{I}_{ca}} 1(\hat{R}_i \leq r)$ and $G_{n_{ca}}(r) = \frac{1}{n_{ca}} \sum_{i \in \mathcal{I}_{ca}} 1(R_i \leq r)$, $r \in \mathbb{R}$. Notice that the lower quantiles $q^{ca}_{\alpha/2}(\hat{R}_i)$ satisfy that $G_{n_{ca}}\{q^{ca}_{\alpha/2}(\hat{R}_i)\} = \frac{n_{ca}+1}{n_{ca}}\alpha/2$ and similarly for the upper quantiles. For the ease of exposition, we write the "lim" as $o_p(1)$ in arguments whenever applicable. Recall that $R(\eta^*) = \eta^*(x) - \eta_{tr}(w)$. It suffices to prove

$$\begin{aligned} 1 - \alpha - o_{n_2}(1) &\leq P\left\{ q^{ca}_{\alpha/2}(\hat{R}_i) \leq R(\eta^*) \leq q^{ca}_{1-\alpha/2}(\hat{R}_i) \right\} \\ &\leq P\left[\frac{(n_{ca}+1)}{n_{ca}} \alpha/2 \leq \hat{G}_{n_{ca}}\left\{R(\eta^*)\right\} \leq \frac{n_{ca}+1}{n_{ca}}(1 - \alpha/2) \right]. \end{aligned}$$

It is easy to show that for any r and a sequence δ_n in Condition 1 where $\delta_n \to 0$ as $n_{tr} \to \infty$,

$$\begin{aligned} &|G_{n_{ca}}(r) - \hat{G}_{n_{ca}}(r + \delta_n)| \leq \frac{1}{n_{ca}} \sum_{i \in \mathcal{I}_{ca}} \left|1(R_i \leq r) - 1(\hat{R}_i \leq r + \delta_n)\right| \\ &= \frac{1}{n_{ca}} \sum_{i \in \mathcal{I}_{ca}} 1(R_i \leq r, \hat{R}_i > r + \delta_n) + \frac{1}{n_{ca}} \sum_{i \in \mathcal{I}_{ca}} 1(R_i > r, \hat{R}_i \leq r + \delta_n). \quad (.1) \end{aligned}$$

Given $\mathcal{D}_{tr}$, the limit of the first term in (.1) is of order $o_{n_{tr}}(1)$ due to

Condition 1 because

$$\lim_{n_{ca}\to\infty}\frac{1}{n_{ca}}\sum_{i\in\mathcal{I}_{ca}}1(R_i\le x,\hat{R}_i>x+\delta_n)\le\lim_{n_{ca}\to\infty}\frac{1}{n_{ca}}\sum_{i\in\mathcal{I}_{ca}}1(|R_i-\hat{R}_i|>\delta_n)$$

$$=\mathbb{E}_{\mathcal{D}_{tr},\mathcal{D}_{ca}}\{1(|R_i-\hat{R}_i|>\delta_n)\}.$$

For the second term in (.1), we have

$$\frac{1}{n_{ca}}\sum_{i\in\mathcal{I}_{ca}}1(R_i>x,\hat{R}_i\le x+\delta_n)=P(R_i>x,\hat{R}_i\le x+\delta_n)+o_{n_{ca}}(1)$$

$$=P\big(\hat{R}_i-\delta_n\le x<R_i\big||R_i-\hat{R}_i|>\delta_n\big)P(|R_i-\hat{R}_i|>\delta_n)+$$

$$P(\hat{R}_i-\delta_n\le x<R_i\big||R_i-\hat{R}_i|\le\delta_n)Pr(|R_i-\hat{R}_i|\le\delta_n)+o_{n_{ca}}(1)$$

$$\le P(|R_i-\hat{R}_i|>\delta_n)+Pr(x\le R_i\le x+2\delta_n)\le F_R(x+2\delta_n)-F_R(x)+o_{n_{ca}}(1)$$

$$\le o_{n_{tr}}(1)+o_{n_{ca}}(1),$$

where F_R is the CDF of scores R_i defined in Theorem 2, which is almost surely continuous. This results in the small term $o_{n_{tr}}(1)$ as $\delta_n\to 0$ when $n_{tr}\to\infty$. Moreover, according to Lei et al. (2018), we have $P\big[G_n\{R(\eta^*)\}\le a\big]\le a+o_{n_{ca}}(1)$ for any $a\in[0,1]$. It follows immediately that

$$P\big\{\hat{G}_n(R_{new})\le\alpha\big\}=P\Big[G_n(R_{new}-\delta_n)\le\alpha+\big\{G_n(R_{new}-\delta_n)-\hat{G}_n(R_{new})\big\}\Big]$$

$$\le P\big\{G_n(R_{new}-\delta_n)\le\alpha+o_{n_{tr}}(1)+o_{n_{ca}}(1)\big\}$$

$$\le\alpha+o_{n_2}(1).$$

Similarly, we can prove the other side of the inequality. Thus, the proof of Lemma 1 is completed. □

Lemma 2. *Refer to the definition of $CI(x_j, \alpha)$ in Algorithm 2. If Condition 1 holds, then for $\alpha \in (0, 0.5)$ and any $(y, x) \in \mathcal{D}_{tst,1}$,*

$$\lim_{n_3 \to \infty} P\left\{\eta^*(x) \in CI(x, \alpha) \mid y = 1\right\} \geq 1 - \alpha.$$

Proof of Lemma 2. Similar to the proof of Lemma 1, we consider two empirical CDFs $\hat{G}_{n_{ca},1}(r) = \frac{1}{n_{ca,1}} \sum_{i \in \mathcal{I}_{ca,1}} 1(\hat{R}_i \leq r)$ and $G_{n_{ca},1}(r) = \frac{1}{n_{ca,1}} \sum_{i \in \mathcal{I}_{ca,1}} 1(R_i \leq r)$, $r \in \mathbb{R}$. Using similar arguments, we can prove that $|G_{n_{ca},1}(r) - \hat{G}_{n_{ca},1}(r + \delta_n)| = o_{n_3}(1)$. Under the assumption that among units with labels being 1, data in $\mathcal{D}_{tst,1}$ and $\mathcal{D}_{ca,1}$ are i.i.d., we have that $P\big[G_{n_{ca},1}\{R(\eta^*)\} \leq a \mid y = 1\big] \leq a + o_{n_{ca,1}}(1)$ for any $a \in [0, 1]$. Following the steps in the last inequality in the proof of Lemma 1, we complete the proof of Lemma 2.

□

Lemma 3. *If Conditions 1 and 2 hold, then for any $\alpha \in (0, 1)$, we have that $|q^{ca,1}_{\alpha/2}(\hat{R}_i) - q^{ca,1}_{\alpha/2}(R_i)| = o_{p,n_1}(1)$. Here a random variable X is said to be of order $o_{p,n}$ if $\lim_{n \to \infty} P(X > \epsilon) = 0$ for any $\epsilon > 0$.*

Proof of Lemma 3. The proof of Lemma 2 establishes that $|G_{n_{ca},1}(r) - \hat{G}_{n_{ca},1}(r + \delta_n)| = o_{n_3}(1)$. The definition of quantiles implies that $\hat{G}_{n_{ca},1}\{q^{ca,1}_{\alpha/2}(\hat{R}_i)\} = \frac{n_{ca,1}+1}{n_{ca,1}} \alpha/2 = G_{n_{ca},1}\{q^{ca,1}_{\alpha/2}(R_i)\}$. It follows from Condition 2 that $\lim_{M \to \infty} P(R_i > M) \to 0$. Thus, the distance between two adjacent jump points, $R_{(s+1)} - R_{(s)}$, is of order $o_{p,n_{ca,1}}(1)$. Furthermore, since both $\hat{G}_{n_{ca}}(\cdot)$ and $G_{n_{ca}}(\cdot)$ are non-decreasing, Lemma 3 follows.

□

Proof of Theorem 1. Applying the Bonferroni Inequality, we have that

$$P\big\{Sens_0(c) \notin CI_{cf}(c,\alpha)\big\} \le P(\mathcal{A}_1) + P(\mathcal{A}_2), \tag{.2}$$

where the two events are $\mathcal{A}_1 = \big\{\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{lo}(\alpha) > c\} > \frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j) > c\}\big\}$ and $\mathcal{A}_2 = \big\{\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j) > c\} > \frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{up}(\alpha) > c\}\big\}$. As the two events are symmetric, $P(\mathcal{A}_1)$ and $P(\mathcal{A}_2)$ will have the same lower bound. Let us focus on the former below. Given $\mathcal{D}_{tr}$, we have

$$\sum_{j\in\mathcal{I}_{tst,1}} 1\big\{b_j^{lo}(\alpha) > c\big\} = \sum_{j\in\mathcal{I}_{tst,1}} 1\big\{w_j^T\beta_{tr} + q_{\alpha/2}^{ca,1}(\hat{R}_i) > c\big\}$$

$$= \sum_{j\in\mathcal{I}_{tst,1}} 1\big\{\eta^*(x_j) - R_j + q_{\alpha/2}^{ca,1}(\hat{R}_i) > c\big\}1\big\{\eta^*(x_j) > c\big\} + 1\big\{w_j^T\beta_{tr} + q_{\alpha/2}^{ca,1}(\hat{R}_i) > c\big\}1\big\{\eta^*(x_j) \le c\big\}$$

$$\le \sum_{j\in\mathcal{I}_{tst,1}} 1\big\{R_j < q_{\alpha/2}^{ca,1}(\hat{R}_i)\big\}1\big\{\eta^*(x_j) > c\big\} + 1\big\{w_j^T\beta_{tr} + q_{\alpha/2}^{ca,1}(\hat{R}_i) > c\big\}1\big\{\eta^*(x_j) \le c\big\}$$

$$\le \sum_{j\in\mathcal{I}_{tst,1}} 1\big\{R_j < q_{\alpha/2}^{ca,1}(\hat{R}_i)\big\} + 1\big\{\eta^*(x_j) > c + R_j - q_{\alpha/2}^{ca,1}(\hat{R}_i)\big\}1\big\{\eta^*(x_j) \le c\big\}.$$

From Lemma 3, we have $|q_{\alpha/2}^{ca,1}(\hat{R}_i) - q_{\alpha/2}^{ca,1}(R_i)| = o_{p,n_1}(1)$. Since $\mathcal{D}_{tst,1}$ and $\mathcal{D}_{ca,1}$ are i.i.d., we get that $|q_{\alpha/2}^{ca,1}(R_i) - q_{\alpha/2}^{tst,1}(R_i)| = o_{p,\min\{n_{ca,1},n_{tst,1}\}}(1)$ (Rio et al., 2017). Thus, $|q_{\alpha/2}^{ca,1}(\hat{R}_i) - q_{\alpha/2}^{tst,1}(R_i)| = o_{p,n_1}(1)$.

Suppose c is randomly chosen from the set of the jump points $\{\eta^*(x_s)\}_{s\in\mathcal{I}_{tst,1}}$.

Then we have

$$\begin{aligned}
&P\left[\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{lo}(\alpha)>c\} > \frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j)>c\}\right]\\
&\le \mathbb{E}\frac{1}{n_{tst,1}}\sum_{s\in\mathcal{I}_{tst,1}} 1\Bigg[\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j)>\eta^*(x_s)\} < \sum_{j\in\mathcal{I}_{tst,1}} 1\{R_j < q^{tst,1}_{\alpha/2}(R_i)+o_{p,n_1}(1)\}+\\
&\qquad\qquad 1\{\eta^*(x_j)>\eta^*(x_s)+R_j-q^{tst,1}_{\alpha/2}(R_i)+o_{p,n_1}(1)\}1\{\eta^*(x_j)\le\eta^*(x_s)\}\Bigg]\\
&\le \mathbb{E}\frac{1}{n_{tst,1}}\sum_{s\in\mathcal{I}_{tst,1}} 1\Bigg[1-q_s<\alpha/2+\frac{1}{n_{tst,1}}\sum_{j:R_j<q^{tst,1}_{\alpha/2}(R_i)} 1\{\eta^*(x_j)\le\eta^*(x_s)\}\Bigg]+o_{n_1}(1)\\
&= \mathbb{E}\frac{1}{n_{tst,1}}\sum_{s\in\mathcal{I}_{tst,1}} 1\big(q_s>1-\alpha\big)+o_{n_1}\le\alpha+o_{n_1}(1),
\end{aligned}$$

where q_s denotes the quantile of $\eta^*(x_s)$ among $\{\eta^*(x_i)\}_{i\in\mathcal{I}_{tst,1}}$. Here, in the first inequality, we can move $o_{p,n_1}(1)$ out of the expectation because of the continuity of probability measure.

Next, we consider the case of c being any cut-off point on $\mathbb{R}$. In this case, there exist two jump points $\eta^*_{(s)}, \eta^*_{(s+1)}$ such that $\eta^*_{(s)}\le c<\eta^*_{(s+1)}$. It follows that

$$
\begin{aligned}
&P\left[\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{lo}(\alpha)>c\} > \frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j)>c\}\right]\\
&\leq \mathbb{E}\frac{1}{n_{tst,1}}\sum_{s\in\mathcal{I}_{tst,1}} 1\left[\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{lo}(\alpha)>\eta^*_{(s)}\} > \frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j)>\eta^*_{(s+1)}\}\right]\\
&\leq \mathbb{E}\frac{1}{n_{tst,1}}\sum_{s\in\mathcal{I}_{tst,1}} 1\left[\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{b_j^{lo}(\alpha)>\eta^*_{(s)}\} > \frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}} 1\{\eta^*(x_j)>\eta^*_{(s)}\}-\frac{1}{n_{tst,1}}\right]\\
&\leq \mathbb{E}\frac{1}{n_{tst,1}}\sum_{s\in\mathcal{I}_{tst,1}} 1\left[1-q_s-\frac{1}{n_{tst,1}}<\alpha/2+\frac{1}{n_{tst,1}}\sum_{j:R_j<q^{tst,1}_{\alpha/2}(R_i)} 1\{\eta^*(x_j)\leq\eta^*(x_s)\}\right]+o_{n_1}(1)\\
&\leq \alpha+o_{n_1}(1).
\end{aligned}
$$

Similarly, we can establish that $P(\mathcal{A}_2)\leq\alpha+o_{n_1}(1)$. Thus, we prove the Theorem 1.

□

Proof of Theorem 2. First, from the consistency of the empirical process (Rio et al., 2017), we know that $Sens(c)=\hat{Sens}(c)+o_{n_{tst}}(1)$. We write $\{\hat{Sens}(c)\notin CI_{cf}(c,\alpha)\}=\mathcal{B}_1\cup\mathcal{B}_2$, with $\mathcal{B}_1=\{\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}}1\{b_j^{lo}(\alpha)>c\}>\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}}1\{\eta_{tr}(w_j)>c\}\}$ and $\mathcal{B}_2=\{\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}}1\{\eta_{tr}(w_j)>c\}>\frac{1}{n_{tst,1}}\sum_{j\in\mathcal{I}_{tst,1}}1\{b_j^{up}(\alpha)>c\}\}$. By the classical theory of the law of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0127

large number, we obtain

$$\frac{1}{n_{tst,1}} \sum_{j \in \mathcal{I}_{tst,1}} \left[1\left\{b_j^{lo}(\alpha) > c\right\} - 1\left\{\eta_{tr}(w_j) > c\right\}\right] \leq \frac{1}{n_{tst,1}} \sum_{j \in \mathcal{I}_{tst,1}} 1\left\{b_j^{lo}(\alpha) > \eta_{tr}(w_j)\right\}$$

$$= P\left\{b_j^{lo}(\alpha) > \eta_{tr}(w_j) | y_j = 1\right\} + o_{n_{tst,1}}(1)$$

$$= P\left\{q_{\alpha/2}^{ca,1}(\hat{R}_i) > 0\right\} + o_{n_{tst,1}}(1).$$

From Lemma 3, we know that $|q_{\alpha/2}^{ca,1}(\hat{R}_i) - q_{\alpha/2}^{ca,1}(R_i)| = o_{p,n_1}(1)$. Also, the consistency of the empirical CDF implies that $|q_{\alpha/2}^{ca,1}(R_i) - F_{R,1}^{-1}(\alpha/2)| = o_{p,n_{ca,1}}(1)$. Under the assumption $P\left\{F_{R,1}^{-1}(\alpha/2) > 0\right\} = o_{n_1}(1)$, we yield that $P\left\{q_{\alpha/2}^{ca,1}(\hat{R}_i) > 0\right\} \leq o_{n_1}(1)$. This completes the proof of Theorem 2.

□

References

Adler, W. and B. Lausen (2009). Bootstrap estimated true and false positive rates and roc curve. *Computational statistics & data analysis 53*(3), 718–729.

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in medicine 13*(5-7), 499–508.

Carel, J.-C. and J. Léger (2008). Precocious puberty. *New England Journal of Medicine 358*(22), 2366–2377.

DiCiccio, T. J. and B. Efron (1996). Bootstrap confidence intervals. *Statistical science 11*(3), 189–228.

Emmanuel, M. and B. R. Bokor (2017). Tanner stages. *London: StatPearls Publishing*.

Euling, S. Y., S. G. Selevan, O. H. Pescovitz, and N. E. Skakkebaek (2008). Role of environmental factors in the timing of puberty. *Pediatrics 121*(Supplement_3), S167–S171.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters 27*(8), 861–874.

Hilgers, R. (1991). Distribution-free confidence bounds for roc curves. *Methods of information in medicine 30*(02), 96–101.

Horvath, S. and K. Raj (2018). Dna methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature reviews genetics 19*(6), 371–384.

Jensen, K., H.-H. Müller, and H. Schäfer (2000). Regional confidence bands for roc curves. *Statistics in medicine 19*(4), 493–509.

Lee, Y., E. T. Tchetgen, and E. Dobriban (2024). Batch predictive inference. *arXiv preprint arXiv:2409.13990*.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association 113*(523), 1094–1111.

Liu, B., Y. Wei, Y. Zhang, and Q. Yang (2017). Deep neural networks for high dimension, low sample size data. In *IJCAI*, Volume 2017, pp. 2287–2293.

Mack, Y. and M. Rosenblatt (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis 9*(1), 1–15.

Marshall, W. A. and J. M. Tanner (1970). Variations in the pattern of pubertal changes in boys. *Archives of disease in childhood 45*(239), 13–23.

McEwen, L. M., K. J. O'Donnell, M. G. McGill, R. D. Edgar, M. J. Jones, J. L. MacIsaac, D. T. S. Lin, K. Ramadori, A. Morin, N. Gladish, et al. (2020). The pedbe clock accurately estimates dna methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences 117*(38), 23329–23335.

Nakas, C. T., L. E. Bantis, and C. A. Gatsonis (2023). *ROC analysis for classification and prediction in practice*. Chapman and Hall/CRC.

Perng, W., M. Tamayo-Ortiz, L. Tang, B. N. Sánchez, A. Cantoral, J. D. Meeker, D. C. Dolinoy, E. F. Roberts, E. A. Martinez-Mier, H. Lamadrid-Figueroa, et al. (2019). Early life exposure in mexico to environmental toxicants (element) project. *BMJ open 9*(8), e030427.

Rio, E. et al. (2017). *Asymptotic theory of weakly dependent random processes*, Volume 80. Springer.

Schafer, H. (1994). Efficient confidence bounds for roc curves. *Statistics in medicine 13*(15), 1551–1561.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world*, Volume 29. Springer.

Xie, M. and Z. Zheng (2022). Homeostasis phenomenon in conformal prediction and predictive distribution functions. *International Journal of Approximate Reasoning 141*, 131–145.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* *3*(1), 32–35.

Department of Biostatistics, University of Michigan

zszheng@umich.edu

Department of Biostatistics, University of Michigan

ybb@umich.edu

Department of Biostatistics, University of Michigan

pxsong@umich.edu