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

Department of Biostatistics, University of Michigan

Supplementary Material

S1 Theoretical Proofs

S1.1 Proof of Proposition 1

As $\eta^*(x) = x^T\beta^* \sim N(0, \beta^{*T}\beta^*)$, $P(y=1) = \mathbb{E}\frac{\exp\{\eta^*(x)\}}{1+\exp\{\eta^*(x)\}} = \frac{1}{2}$. Then,

$$OR(c;w) > 1 \Leftrightarrow \big|\frac{1}{4} - P\big\{\eta_{tr,1}(w) > c, y = 1\big\}\big| > \big|\frac{1}{4} - P\big\{\eta_{tr,2}(w) > c, y = 1\big\}\big|.$$

Notice that since β_{tr} is a consistent estimator, $\eta_{tr,1} - \eta^*(x) = x^T(\beta_{tr} - \beta^*) = o_{n_{tr},p}(1)$. Denoting the density function of $\eta^*(x)$ as $f_{\eta^*}(x)$, we have that

$$P\big\{y = 1, \eta_{tr,1}(w) > c\big\} = \int_{-\infty}^{\infty} P\big\{y = 1, \eta_{tr,1}(w) > c \mid \eta^*(x) = t\big\} f_{\eta^*}(t)dt$$

$$= o_{n_{tr}}(1) + \int_{-\infty}^{\infty} P\big\{y = 1 \mid \eta^*(x) = t\big\} f_{\eta^*}(t)1(t > c)dt = o_{n_{tr}}(1) + g_1(c),$$

and

$$P\big\{y=1,\eta_{tr,2}(w)>c\big\}=\int_{-\infty}^{\infty}P\big\{y=1,\eta_{tr,2}(w)>c\mid\eta^*(x)=t\big\}f_{\eta^*}(t)dt$$
$$=\int_{-\infty}^{\infty}P\big\{y=1\mid\eta^*(x)=t\big\}f_{\eta^*}(t)P\{\eta_{tr,2}(w)>c\mid\eta^*(x)=t\}dt=g_2(c).$$

We can see that both probabilities can be written as continuous decreasing functions $g_1(c)$ and $g_2(c)$ in $c\in\mathbb{R}$, with support inside $[0,\frac{1}{2}]$. We further notice that at the point c^* such that $g_2(c^*)=\frac{1}{4}$, we have $g_1(c^*)>\frac{1}{4}$. Thus there exists an interval $C\in[0,\frac{1}{2}]$ around c^*, such that for any $c\in C$, $|\frac{1}{4}-g_1(c)|>|\frac{1}{4}-g_2(c)|$, thus proves the proposition.

S1.2 Proof of Proposition 2

From Lemma B3 in the Appendix, we know that for lower quantiles $q^{ca,1}_{\alpha/2}(\hat{R}_{i,1})=q^{ca,1}_{\alpha/2}(R_{i,1})+o_{p,n_1}(1)$ and $q^{ca,1}_{\alpha/2}(\hat{R}_{i,2})=q^{ca,1}_{\alpha/2}(R_{i,2})+o_{p,n_1}(1)$. The same results hold for the upper quantile of $1-\alpha/2$. Notice that $R_{i,1}=\eta^*(x)-\eta_{tr,1}(w)=x^T(\beta^*-\beta_{tr})=o_{p,n_{tr}}(1)$, and $R_{i,2}=\eta^*(x)-\eta_{tr,2}(w)=x^T(\beta^*-\gamma_{tr})+x_1\gamma_1\sim N(0,\sigma_2^2)$, where $\sigma_2^2=\beta_1^{*2}+(\beta^*_{-1}-\gamma_{-1,tr})^T(\beta^*_{-1}-\gamma_{-1,tr})>0$. Following the proof in Xie & Zheng (2022), we can show that $q^{ca,1}_{\alpha/2}(\hat{R}_{i,2})=o_{p,n_1}+\sigma_2\Phi(\alpha/2)>o_{p,n_1}=q^{ca,1}_{\alpha/2}(\hat{R}_{i,1})$, where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution. Similarly, we can prove the inequality for the upper quantile $1-\alpha/2$. Combining them, we prove the proposition.

S1.3 Proof of Proposition 3

Noticing that the upper and lower limits of the confidence interval are symmetric, we only need to show $Sens_0(c^\dagger) + Spec_0(c^\dagger) \leq \hat{Sens}(c^\dagger; b_1^{up}) + \hat{Spec}(c^\dagger; b_0^{lo}) + \alpha$, and the lower bound will follow using similar arguments. For the upper bound, we have that

$$
\begin{aligned}
&Sens_0(c^\dagger) - \hat{Sens}(c^\dagger; b_1^{up}) \\
&= \sum_{j \in \mathcal{I}_{tst,1}} \frac{1\{\eta^*(x_j) > c^\dagger\}}{n_{tst,1}} - \sum_{j \in \mathcal{I}_{tst,1}} \frac{1\{\eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i) > c^\dagger\}}{n_{tst,1}} \\
&= \frac{1}{n_{tst,1}} \left[\sum_{j \in \mathcal{I}_{tst,1}} 1\{\eta^*(x_j) > c^\dagger\} - 1\{\eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i) > c^\dagger\} \right] \\
&= \frac{1}{n_{tst,1}} \Bigg[\sum_{j \in \mathcal{I}_{tst,1}} 1\{\eta^*(x_j) > c^\dagger \geq \eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i)\} \\
&\qquad\qquad - \sum_{j \in \mathcal{I}_{tst,1}} 1\{\eta^* \leq c^\dagger < \eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i)\} \Bigg] \\
&= \frac{1}{n_{tst,1}} \Bigg[\sum_{j \in \mathcal{I}_{tst,1}} 1\{R_j > c^\dagger - \eta_{tr}(w_j) \geq q^{ca,1}_{1-\alpha/2}(\hat{R}_i)\} \\
&\qquad\qquad - \sum_{j \in \mathcal{I}_{tst,1}} 1\{\eta^* \leq c^\dagger < \eta_{tr}(w_j) + q^{ca,1}_{1-\alpha/2}(\hat{R}_i)\} \Bigg] \\
&\leq \frac{1}{n_{tst,1}} \sum_{j \in \mathcal{I}_{tst,1}} 1\{R_j > q^{ca,1}_{1-\alpha/2}(\hat{R}_i)\} = \alpha/2 + o_{n_1}(1).
\end{aligned}
$$

The last equality is reasoned in the proof of Theorem 1 in the Appendix. Then similarly, letting $n_0 = \min\{|\mathcal{D}_{tr}|, |\mathcal{D}_{ca,0}|, |\mathcal{D}_{tst,0}|\}$, we obtain

$$
\begin{aligned}
&Spec_0(c^\dagger) - \hat{Spec}(c^\dagger; b_1^{up}) \\
&= \frac{1}{n_{tst,0}} \Bigg[\sum_{j \in \mathcal{I}_{tst,0}} 1\big\{ R_j \le c^\dagger - \eta_{tr}(w_j) < q_{\alpha/2}^{ca,0}(\hat{R}_i) \big\} \\
&\qquad\qquad - \sum_{j \in \mathcal{I}_{tst,0}} 1\big\{ \eta_{tr}(x_j) + q_{\alpha/2}^{ca,0}(\hat{R}_i) \le c^\dagger < \eta^*(x_j) \big\} \Bigg] \\
&\le \frac{1}{n_{tst,0}} \sum_{j \in \mathcal{I}_{tst,0}} 1\big\{ R_j \le q_{\alpha/2}^{ca,0}(\hat{R}_i) \big\} = \alpha/2 + o_{n_0}(1).
\end{aligned}
$$

Thus, as $\min\{|\mathcal{D}_{tr}|, |\mathcal{D}_{ca}|, |\mathcal{D}_{tst}|\} \to \infty$, we surely have $Sens_0(c^\dagger) + Spec_0(c^\dagger) \le \hat{Sens}(c^\dagger; b_1^{up}) + \hat{Spec}(c^\dagger; b_0^{lo}) + \alpha$. This completes the proof of the Proposition.

S2 Additional numerical results

S2.1 Additional numerical results in Sections 1 and 2

This section provides additional numerical results for the simulation experiments described in Section 1.2 and Section 2.3.

We generate three data sets of random samples, D_{tr}, D_{ca} and D_{tst} with different sample sizes of 500 and 5000 for each data set. Following the same model, we simulate data from a logistic 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 four classifiers: (TM1) $\eta_{tr,1} = x_1\beta_{1,tr}+x_2\beta_{2,tr}+x_3\beta_{3,tr}$, (TM2) $\eta_{tr,2} = x_2\gamma_{2,tr} + x_3\gamma_{3,tr}$, (TM3) $\eta_{tr,3} = x_1\psi_{1,tr} + x_2\psi_{3,tr}$, and (TM4) 10-layer Neural Network model with outcome y and covariates x_1, x_2, x_3. Following the standard model training procedure using the R functions `glm` and `neuralnet`, we estimate the parameters $\beta_{tr}, \gamma_{tr}, \psi_{tr}$ using $\mathcal{D}_{tr} \cup \mathcal{D}_{ca}$ for the standard bootstrap method and $\mathcal{D}_{tr}$ for the proposed conformal method.

In the application of the bootstrap method, in addition to the standard version introduced in Section 1.2, we include an additional bootstrap method, referred to as "double bootstrap in this paper, which is operated under a parametric bootstrap method for modeling training $\eta_{tr}(w)$ and a non-parametric bootstrap method for the empirical sensitivity $Sens(c)$. In the implementation of "double bootstrap", we first obtain resampling copies of $\eta_{tr}^{b_1}$, and then for individual copies of $\eta_{tr}^{b_1}$, we calculate its resampled $Sens(c)^{b_2}$, resulting in a construction of the confidence interval. The details are given in the following Algorithm.

Algorithm S1. Constructing a bootstrap confidence interval for $Sens(c)$ with a given $c \in \mathbb{R}$.

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

Step II: For $b_1 \in \{1, 2, \cdots, B\}$, run the following loops.

Step II.a Draw bootstrap sample $\mathcal{D}_{tr}^{b_1} = \{(y_j^{b_1}, x_j) : y_j^{b_1} \sim Ber\big(\pi_{tr}(w_j)\big)\}$.

Step II.b Fit the logistic training model $\eta_{tr}^{b_1}$ on the bootstrap data $\mathcal{D}_{tr}^{b_1}$.

Step II.c: Given b_1, for $b_2 \in \{1, 2, \cdots, B\}$, run the following loops.

Step II.c1: Draw bootstrap sample $\mathcal{D}_{tst}^{b_2} = \{(y_j^{b_2}, x_j^{b_2}) : j \in \mathcal{I}_{tst}^{b_2}\}$ and set $\mathcal{I}_{tst,1}^{b_2} = \{j \in \mathcal{I}_{tst} : y_j^{b_2} = 1\}$.

Step II.c2: Calculate $\hat{Sens}^{b_1,b_2}(c) = \frac{1}{n_{tst,1}} \sum_{j \in \mathcal{I}_{tst,1}^{b_2}} 1\{\eta_{tr}^{b_1}(w_j) > c\}$ where $n_{tst,1} = |\mathcal{I}_{tst,1}^{b_2}| = |\mathcal{I}_{tst,1}|$.

Step III: Compute the empirical CDF

$$\hat{F}_B(t) = \frac{1}{B^2} \sum_{b_1,b_2=1}^{B} 1\big[\sqrt{n_{tst,1}}\big\{\hat{Sens}^{b_1,b_2}(c) - \hat{Sens}(c)\big\} \le t\big].$$

Step IV: Output $CI_B(c;\alpha)$ of the form:

$$\big[\hat{Sens}(c) - \hat{F}_B^{-1}(1-\alpha/2)/\sqrt{n_{tst,1}},\ \hat{Sens}(c) - \hat{F}_B^{-1}(\alpha/2)/\sqrt{n_{tst,1}}\big].$$

In addition to the t-bootstrap method, we also considered two alternative approached to constructing bootstrap intervals from the bootstrap samples generated from Algorithm 1 and Algorithm S1; they are, (a) $CI_{B,\sigma}(c;\alpha) = \big[\hat{Sens}(c) - 1.96\hat{\sigma}, \hat{Sens}(c) + 1.96\hat{\sigma}\big]$ and (b) $CI_{B,emp}(c;\alpha) = \big[q_{\alpha/2}\{\hat{Sens}^b(c)\}, q_{1-\alpha/2}\{\hat{Sens}^b(c)\}\big]$, where $q_\gamma(\hat{Sens}^b(c))$ denote the γ-th quantile of $\{\hat{Sens}^b(c)\}_b^B$.

To compare the proposed conformal method, we consider the batch conformal inference method developed in Lee et al. (2024) in a comparison. To reduce the randomness from the simulated data set, we repeatedly generate the data sets $\mathcal{D}_{tr}, \mathcal{D}_{ca}$ and $\mathcal{D}_{tst}$ for 50 times and repeatedly imple-

ment both bootstrap and conformal methods for each of the data set. All the results are given in the following Tables S1 - S3. The main results for the average coverage rate and length of the confidence interval are reported in Table S1. The results confirmed the theoretical properties presented in the main text: (a) The conformal confidence bands have an adequate coverage (around 95% nominal level) for the oracle $Sens_0$ and $spec_0$; (b) the non-parametric bootstrap confidence bands guarantee a proper coverage for $Sens_0$ and $Spec_0$ only when the working model is unbiased; (c) the double bootstrap method also shows proper coverage, but with slightly larger length, and the batch conformal inference method has similar results as those given by our conformal method.

In Table S2, we show that our proposed method achieves the best computational efficiency; even with a large sample size of $|\mathcal{D}_{tr}| = |\mathcal{D}_{ca}| = |\mathcal{D}_{tst}| = 5000$, our proposed method can easily finish the calculations within minutes. In contrast, the double bootstrap method and the batch conformal inference method take much longer time in their calculations.

Finally, to explain the bad coverage for non-parametric bootstrap method, we present Table S3 to show that this method only tries to cover $Sens(c)$ but not the real target $Sens_0(c)$, leading to under-estimated uncertainty, which is problematic in practice.

		coverage for $Sens_0(c)$				average length			
		(TM1)	(TM2)	(TM3)	(TM4)	(TM1)	(TM2)	(TM3)	(TM4)
non-parametric bootstrap (B=1000)	t-bootstrap	0.933	0.734	0.324	0.951	0.23	0.22	0.25	0.23
	2-sigma bootstrap	0.932	0.503	0.335	0.952	0.22	0.22	0.24	0.23
	empirical bootstrap	0.943	0.689	0.325	0.945	0.22	0.23	0.25	0.22
double bootstrap ($B_1 = B_2 = 500$)	t-bootstrap	0.986	0.962	0.984	0.961	0.35	0.59	0.81	0.31
	2-sigma bootstrap	0.965	0.975	0.973	0.973	0.35	0.62	0.82	0.35
	empirical bootstrap	0.978	0.965	0.981	0.982	0.34	0.63	0.79	0.31
conformal method	batch conformal inference	0.952	0.952	0.932	0.943	0.31	0.56	0.78	0.28
	our proposed method	0.945	0.941	0.935	0.956	0.33	0.55	0.75	0.31

Table S1: Comparisons on coverage rate and average interval length across two bootstrap methods (non-parametric and double bootstrap), three interval construction approaches, and two conformal methods among four classifiers (TM1)-(TM4). "B" stands for the bootstrap sample size, and 50 rounds of simulations are conducted to draw the summary statistics.

S2.2 Additional numerical results in Section 3

In this section, we first provide the averaged confidence intervals for AUC of specificity using conformal method: the 80% CI is $[0.462, 0.890]$, the 90% CI is $[0.384, 0.921]$ and the 95% CI is $[0.246, 0.925]$. As discussed in the main text, all three confidence intervals for specificity contain 0.5, showing the existence of large uncertainty in the prediction of partial maturation.

Next, we provide some supplementary figures to the plots in Section 3 of the main text at higher confidence levels. Figures S1 and S2 correspond

		$n_{tr} = n_{ca} = n_{tst} = 500$				$n_{tr} = n_{ca} = n_{tst} = 5000$			
		TM1	TM2	TM3	TM4	TM1	TM2	TM3	TM4
non-parametric bootstrap (B=1000)	t-bootstrap	31s	30s	27s	123s	73s	68s	62s	933s
	2-sigma bootstrap	30s	31s	28s	129s	70s	64s	60s	930s
	empirical bootstrap	28s	26s	25s	121s	62s	59s	58s	921s
double bootstrap ($B_1 = B_2 = 500$)	t-bootstrap	423s	412s	392s	874s	1623s	1587s	1566s	$\geq$1hr
	2-sigma bootstrap	427s	410s	402s	862s	1634s	1601s	1543s	$\geq$1hr
	empirical bootstrap	398s	388s	374s	802s	1602s	1520s	1459s	$\geq$1hr
conformal method	batch conformal inference	623s	612s	589s	1211s	$\geq$1hr	$\geq$1hr	$\geq$1hr	$\geq$1hr
	our proposed method	6s	5s	5s	30s	25s	21s	20s	198s

Table S2: Comparison of run-time (in seconds) for one round of simulation under two bootstrap methods (non-parametric and double bootstrap), three interval construction approaches, and two conformal methods among four classifiers (TM1)-(TM4). "B" stands for the bootstrap sample size, and 50 rounds of simulations are conducted to draw the summary statistics.

to the 90% and 95% confidence bands for sensitivity and specificity, respectively. We can see from these two figures that the 90% and 95% confidence bands are wider and only 90% confidence interval for AUC of sensitivity is above 0.5. This indicates that the prediction model exhibits a large uncertainty at the confidence level 90% and 95%.

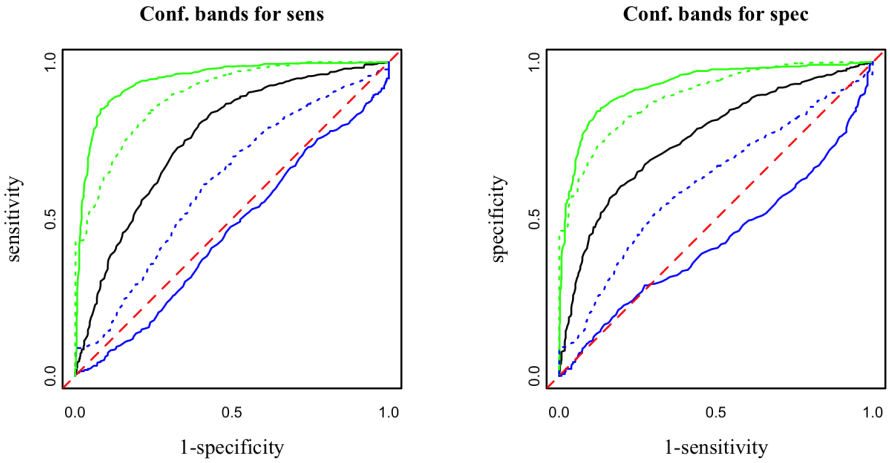

Figure S1: The 90% ROC confidence bands for sensitivity (left) and specificity (right) produced by conformal and bootstrap methods. Black solid line is the ROC of the sexual maturation prediction; the green solid and dashed line correspond to the upper confidence bands, while the blue solid and dashed line are the lower confidence bands of the conformal and bootstrap methods, respectively. The red dashed line is the reference line of random decision.

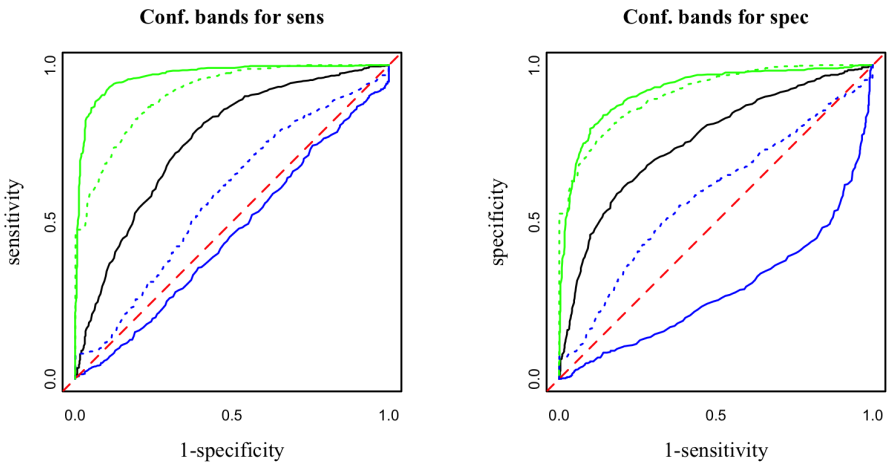

Figure S2: The 95% ROC confidence bands for sensitivity (left) and specificity (right) produced by conformal and bootstrap methods. Black solid line is the ROC of the sexual maturation prediction; the green solid and dashed line correspond to the upper confidence bands, while the blue solid and dashed line are the lower confidence bands of the conformal and bootstrap methods, respectively. The red dashed line is the reference line of random decision.

		coverage for $Sens(c)$			
		(TM1)	(TM2)	(TM3)	(TM4)
non-parametric bootstrap (B=1000)	t-bootstrap	0.962	0.953	0.954	0.993
	2-sigma bootstrap	0.974	0.945	0.945	0.982
	empirical bootstrap	0.959	0.952	0.962	0.987
double bootstrap ($B_1 = B_2 = 500$)	t-bootstrap	1	1	0.992	1
	2-sigma bootstrap	1	0.995	0.995	1
	empirical bootstrap	0.991	0.993	0.986	0.999
conformal method	batch conformal inference	0.996	0.991	0.982	1
	our proposed method	1	1	0.995	1

Table S3: Comparisons on Coverage for the empirical sensitivity $Sens(c)$ across two bootstrap methods (non-parametric and double bootstrap), three interval construction approaches, and two conformal methods among four classifiers (TM1)-(TM4). "B" stands for the bootstrap sample size, and 50 rounds of simulations are conducted to draw the summary statistics.

Bibliography

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

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