Conformal Prediction and Recidivism: Take 2

Chris Banerji

May 2023

1 The importance of personalised prediction

Here we consider data collected by ProPublica in 2016 (`https://github.com/propublica/compas-analysis`) describing predictions of recidivism rates by the COMPAS algorithm (used in USA courts in a number of states), alongside recorded re-offences in the 2 years following COMPAS prediction. Data was collected for 3607 African-American individuals and 2398 Caucasian individuals. Analysis of this data showed that while COMPAS has comparable accuracy on the two ethnic groups, the distribution of inaccuracies is troubling, with African-Americans more likely to predicted to re-offend when they do not, and Caucasians more likely to be predicted not to re-offend when they do. This finding motivates an investigation of personalised prediction uncertainty in this data, to identify individuals for whom COMPAS predictions can be relied upon in decision making and importantly, those for whom it can not.

Given that predictions and outcomes are known, but the methodology by which COMPAS makes predictions is proprietary and not in the public domain, we consider Conformal Prediction to understand uncertainty in this data.

The ProPublica data set provides the raw COMPAS scores, whether each individual re-offended in the 2 years after the score was calculated and if so, time to re-offence. A Cox-proportional hazards model confirms that the COMPAS score is significantly positively associated with time to re-offence across the whole data set (HR$= 1.68(1.61, 1.75)$, Wald $p < 2.2 \times 10^{-16}$). Conformalised survival analysis is still in its infancy, but conformal prediction can prove a powerful tool when applied to classification problems. We thus consider the following framework:

We sample a random training set X_{train} of 2000 individuals from the ProPublica data, as well as a disjoint random calibration set X_{calib} of 2000 individuals. The remaining data we use as a test set.

Using the training data we perform survival analysis of the COMPAS score against recidivism probability. Employing the *survminer* and *maxstat* packages in R we derive 2 COMPAS score cut-off values (c_{low} and c_{high}), which optimally divide training data samples into low (COMPAS score $< c_{low}$), medium (COMPAS score $\in [c_{low}, c_{high}]$) and high (COMPAS score $> c_{high}$) recidivism risk groups, based on their survival probability. These 3 risk groups represent our prediction outcomes.

For each individual i in our test set we want to derive a prediction set $C(i)$ which contains the true outcome y_i (low, medium and high re-offence risk) with a specified certainty α. This equates the following coverage guarantee:

$$\mathbb{P}[y_i \in C(i)] \geq 1 - \alpha \tag{1}$$

For each individual in the calibration set we must compute a conformal score which quantifies the inaccuracy of the COMPAS prediction x, relative to a known true re-offence risk class y. To do

this we must first define a 'truth' for our low, medium and high risk categories. We propose that a low risk prediction is true if the individual does not re-offend in 2 years, a high risk prediction is true if the individual re-offends in under 1 year and a medium risk prediction is true if the individual re-offends in between 1 and 2 years. We thus define our conformal score as follows:

$$s(x,y) = \begin{cases} 0 & \text{if } x < c_{low} \text{ and } y = \text{no offence in 2 years} \\ abs(x - c_{low}) & \text{if } x > c_{low} \text{ and } y = \text{no offence in 2 years} \\ 0 & \text{if } x \in [c_{low}, c_{high}] \text{ and } y = \text{re-offence in 1-2 years} \\ min\Big[abs(x - c_{low}), abs(x - c_{high})\Big] & \text{if } x \notin [c_{low}, c_{high}] \text{ and } y = \text{re-offence in 1-2 years} \\ 0 & \text{if } x > c_{high} \text{ and } y = \text{re-offence in } < 1 \text{ year} \\ abs(x - c_{high}) & \text{if } x < c_{high} \text{ and } y = \text{no offence in 2 years} \end{cases} \tag{2}$$

Simply put, the conformal score is 0 if the COMPAS prediction lies in the range of the appropriate risk category and is otherwise the smallest positive distance between the COMPAS prediction and a prediction value which would assign the correct risk category.

Due to the way our conformal score is defined, for each outcome the conformal score distribution is not the same in general. We thus consider an adapted form of conformal prediction described by Angelopoulos and Bates 2022, called class conditional conformal prediction.

Let $\hat{q_\alpha^y}$ be the $\frac{\lceil (n_y+1)(1-\alpha) \rceil}{n^y}$ quantile of $s(\{X_{calib}|Y_{calib} = y\}, y)$ where $n_y = |\{X_{calib}|Y_{calib} = y\}|$. Appealing to class conditional conformal prediction we can derive our prediction sets via:

$$C(X_{test}) = \{y : s(X_{test}, y) < \hat{q_\alpha^y}\}. \tag{3}$$

Under class conditional conformal prediction, the marginal coverage guarantee is slightly stronger and holds conditional on the true outcome.

We performed the above procedure for $\alpha = 0.1$, repeating the training, calibration and test data set re-sampling 30 times. On average the accuracy of the training set derived COMPAS score cut-offs (point-estimates) in assigning test set individuals to the true risk group (low/medium/high) was little better than average (mean accuracy: 39.3% Fig 1A.). By design, however the coverage of the conformal prediction sets - i.e., the proportion of time they contained the true class - was almost exactly 90% (mean coverage: 89.9%, Fig. 1B). The distributions of c_{low} and c_{high} were unimodal across the 30 random training sets (mean $c_{low} = 0.386$, mean $c_{high} = 0.627$ Fig. 1C & D), as were the distributions of the quantile cut-offs for each outcome $\hat{q^y_{alpha}}$ (Fig. 1E-G).

If the prediction set for an individual in the test set contained only one outcome, we labelled that prediction as certain, while prediction sets with multiple outcomes were considered uncertain. While at the $\alpha = 0.1$ none of our three outcomes were found in every prediction set, all prediction sets containing the medium risk outcome were uncertain, while high and low risk containing prediction sets could be either certain or uncertain. Our approach argues that certain predictions are more likely to be accurate while uncertain predictions are less likely to be accurate and should not relied upon. To investigate this we considered the high and low risk groups for which we have both certain and uncertain predictions across our 30 random test set samples. Individuals assigned low risk by point estimate without appealing to uncertainty were truly low risk (i.e., no reoffence in 2 years) on average 69.6% of the time. However, certain low risk predictions were significantly more accurate (Wilcox $p < 2.2 \times 10^{-16}$, mean acc: 79.9%) and uncertain low risk predictions significantly less accurate (Wilcox $p < 6.2 \times 10^{-7}$, mean acc: 64.6%, Fig. 2A). Individuals assigned high risk by point estimate without appealing to uncertainty were truly high risk (i.e., re-offence in 1 year) on average 53.6% of the time. However, certain high risk predictions were significantly more accurate (Wilcox

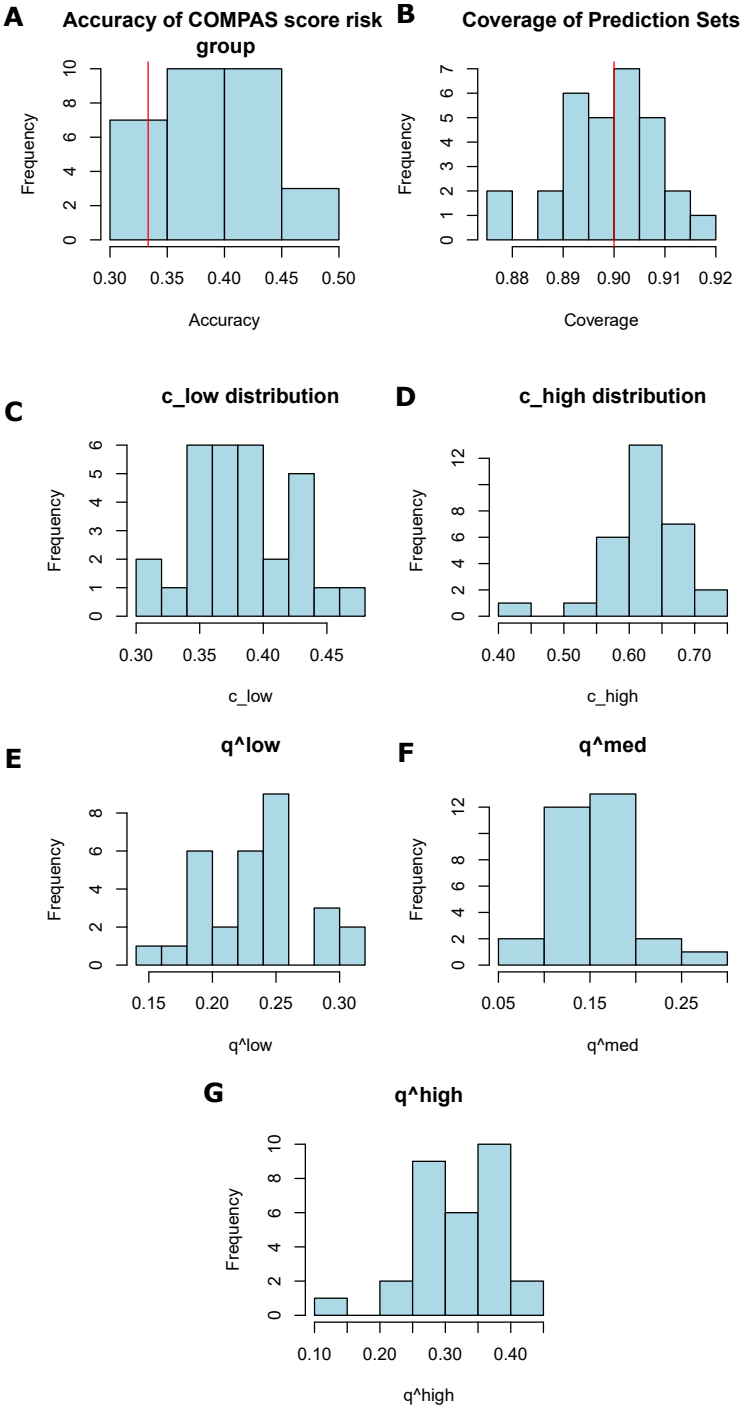

Figure 1: *Conformal Prediction Sets Compared to Accuracy.* A. Histogram displaying the total accuracy of recidivism risk class allocation for 30 random test sets, with COMPAS score cut-offs derived from 30 random training sets. A red line denotes accuracy expected from random allocation to classes (1/3). B. Histogram displaying the coverage of the prediction sets of the same 30 random test sets derived using calibration sets, with $\alpha = 0.1$. A red line denotes the expected coverage of 90% for this choice of α. Histograms display the distributions of C. c_{low}, D. c_{high}, E. q_α^{low}, F. q_α^{med}, G. q_α^{high}. We see reasonable consistency across the random train/calibration/test sets.

$p < 1.2 \times 10^{-5}$, mean acc: 60.9%) and uncertain high risk predictions similarly accurate (Wilcox $p = 0.1$, mean acc: 52.2%, Fig. 2B).

To provide a clearer visual example of the power of including uncertainty in our analysis we considered a representative random train/calb/test set. We performed 2 survival analyses - in the first we grouped patients into risk class based on the COMPAS point-estimate and the cut-offs derived from the training data (Fig 2C), in the second we sub grouped these patients based on uncertainty (Fig 2D). The results are striking - without considering uncertainty 27% of individuals allocated to a low risk class re-offend, however when we only consider certain low risk predictions only 17% re-offend. Conversely, without uncertainty we allocate 35% of individuals allocated to the high risk class do not re-offend, while considering certainty we reduce this proportion to less than half (16%).

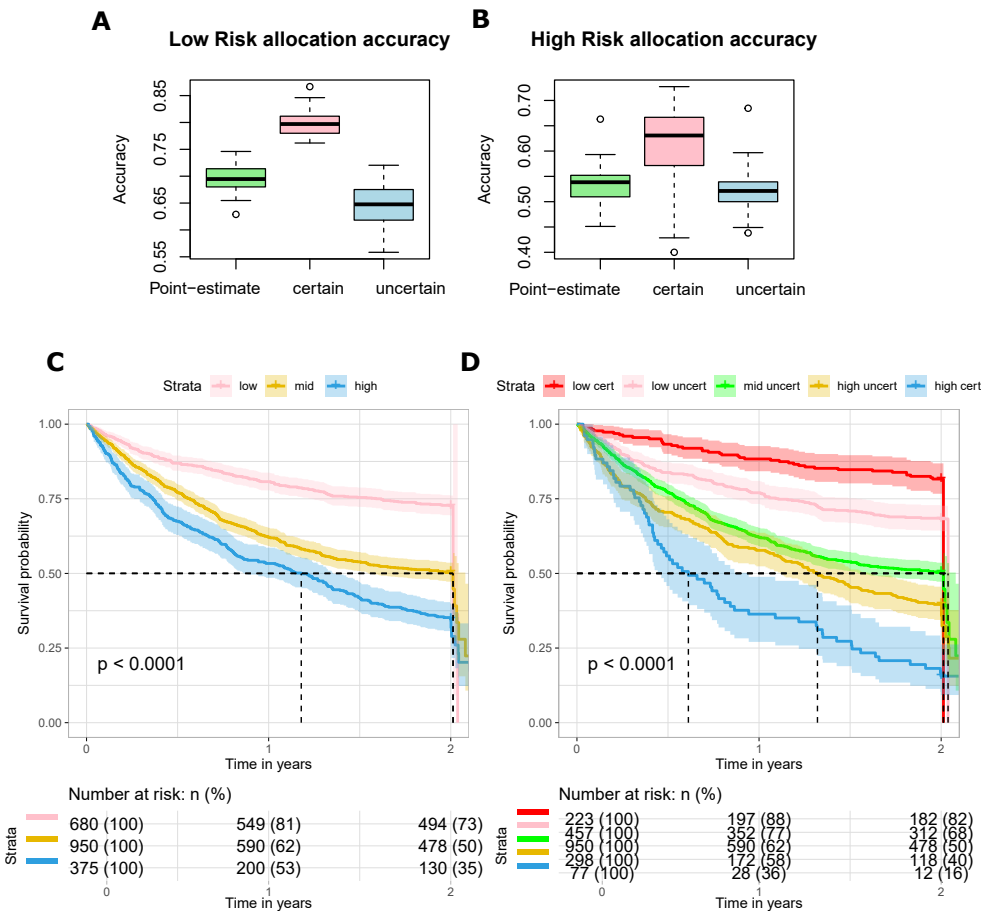

Figure 2: *The importance of personalised uncertainty* Boxplots display the accuracy of point-estimates without uncertainty, certain and uncertain predictions, for A. low and B. high risk class prediction (compared to true re-offence rate), across 30 train/calib/test set randomisations. Kaplan-Meir plots display survival (recidivism) curves for C. point estimates without considering uncertainty and D. estimates considering uncertainty, in a representative train/calib/test set.

The COMPAS algorithm has been criticized for being unfair to African-Americans compared to Caucasians in terms of accuracy of prediction. We have shown that uncertain predictions are less likely to be accurate. Personalised uncertainty in our conformal prediction setting can be computed as the size of the prediction set for an individual in the test set. If the COMPAS score is more inaccurate for African-Americans compared to Caucasians, we would expect personalised uncertainty to be higher in African-Americans. We find that this is indeed the case (Wilcoxon $p < 2.2 \times 10^{-16}$, Fig 3A). To investigate how accuracy, uncertainty and ethnicity interact in the ProPublica data set, we next considered the high and low risk categories for which we have both certain and uncertain predictions.

African-American individuals who were assigned low risk point estimate without uncertainty did not re-offend 67.4% of the time, with this rate rising significantly to 77.9% for certain predictions (paired Wilcoxon $p < 1.9 \times 10^{-9}$, Fig 3B). Caucasian individuals who were assigned low risk point estimate without uncertainty did not re-offend 71.2% of the time, with this rate rising significantly to 80.7% for certain predictions (Wilcoxon $p < 1.9 \times 10^{-9}$, Fig 3C). Including certainty thus led to more accurate low risk predictions for both ethnicities.

African-American individuals who were assigned high risk point estimate without uncertainty re-offended within 1 year 53.9% of the time, with this rate significantly rising to 61.5% for certain predictions (Wilcoxon $p < 1.1 \times 10^{-4}$, Fig 3D). Caucasian individuals who were assigned high risk point estimate without uncertainty re-offended within 1 year 52.4% of the time, with this rate rising only non-significantly to 58.5% for certain predictions (Wilcoxon $p = 0.07$, Fig 3E). Importantly for high risk predictions incorporating certainty resulted in significantly more accurate predictions for African-Americans than for Caucasians. This finding suggests that the COMPAS prediction score is indeed less accurate for African-Americans compared to Caucasians when a high re-offender risk prediction is made. Incorporating uncertainty can help to offset this unfairness.

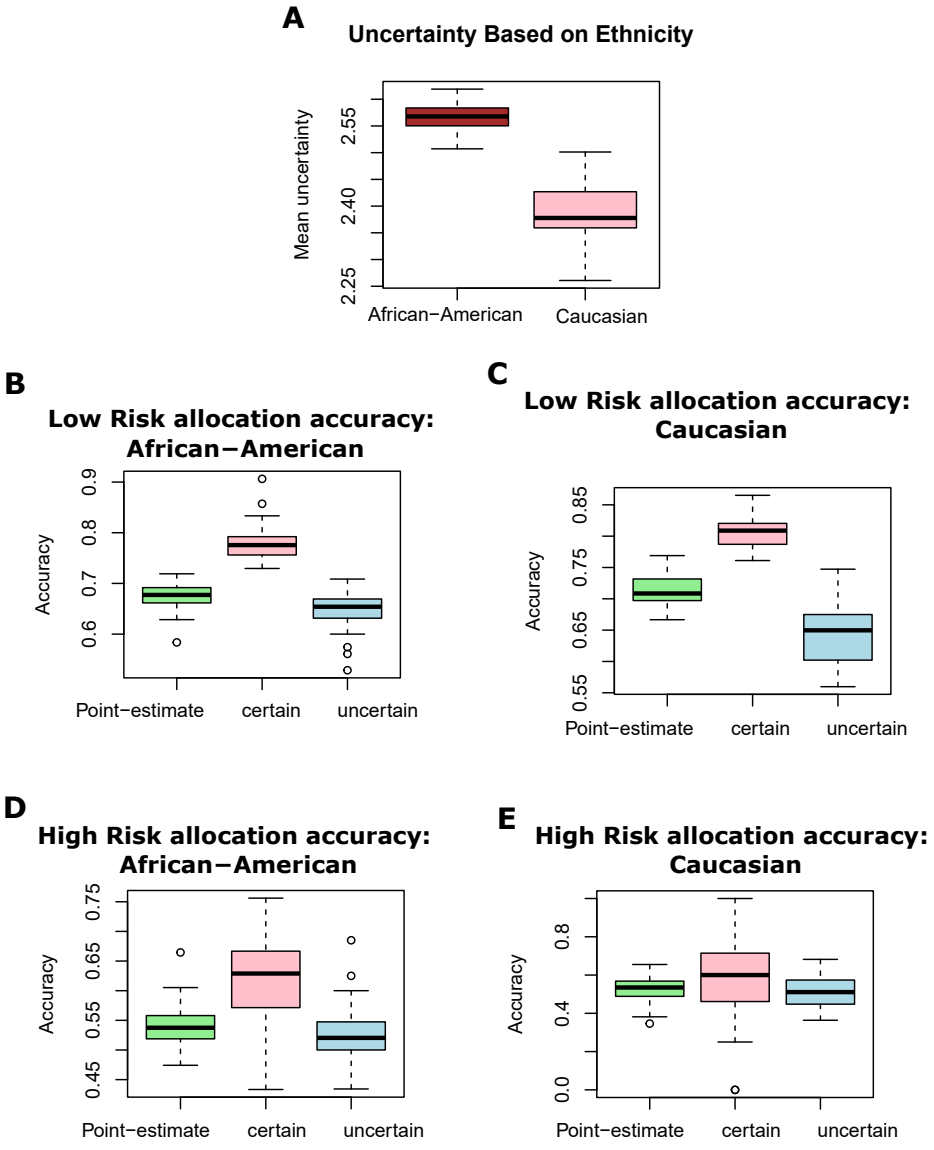

Figure 3: *Personalised uncertainty and protected characteristics* A. Boxplot displays personalised uncertainty (prediction set size) for African-American individuals compared to Caucasian individuals. Boxplots display the accuracy of point-estimates without uncertainty, certain and uncertain predictions, for A. African-American low risk B. Caucasian low risk C. African-American high risk D. Caucasian high risk predictions (compared to true re-offence rate), across 30 train/calib/test set randomisations.