NESTED CONFORMAL PREDICTION SETS FOR CLASSIFICATION WITH APPLICATIONS TO PROBATION DATA

BY ARUN K. KUCHIBHOTLA\textsuperscript{1} AND RICHARD A. BERK\textsuperscript{2}

\textsuperscript{1}Carnegie Mellon University
\textsuperscript{2}University of Pennsylvania

Risk assessments to help inform criminal justice decisions have been used in the United States since the 1920s (Burgess, 1928). Over the past several years, statistical learning risk algorithms have been introduced amid much controversy about fairness, transparency and accuracy. In this paper, we focus on accuracy for a large department of probation and parole that is considering a major revision of its current, statistical learning risk methods. Because the content of each offender’s supervision is substantially shaped by a forecast of subsequent conduct, forecasts have real consequences. Here we consider the probability that risk forecasts are correct. We augment standard statistical learning estimates of forecasting uncertainty (i.e., confusion tables) with uncertainty estimates from nested conformal prediction sets. In a demonstration of concept using data from the department of probation and parole, we show that the standard uncertainty measures and uncertainty measures from nested conformal prediction sets can differ dramatically in formulation and output. We also provide a modification of nested conformal called the localized conformal method to match confusion tables more closely. A strong case can be made favoring the nested conformal approach. As best we can tell, our formulation of such comparisons and consequent recommendations is novel.

1. Introduction. Risk assessments have been used in the United States to help inform criminal justice decisions since the 1920s (Burgess, 1928). Typically, these risk assessments forecast whether a given individual, already convicted of a crime, will re-offend. Over the last decade, statistical learning assessments of risk have been introduced in some jurisdictions, often with considerable controversy. Concerns include fairness, transparency, and accuracy. All three are important, but here we focus on accuracy. Excellent treatments of fairness and transparency can be found elsewhere (Carlson, 2017; Berk, 2019; Coglianese and Lehr, 2019; Huq, 2018; Kearns and Roth, 2019; Rudin et al., 2020).


Keywords and phrases: Conformal Prediction, Level Sets, Classification, Risk Assessment, Criminal Justice, Statistical Learning
In this paper, we analyze proprietary data from a department of probation and parole in a large metropolitan area that has for several years used a statistical learning classifier to assess the risk of re-offending for the probationers being supervised.\(^1\) As described below, the risk procedure has performed well, but a major revision with current data is under consideration. Forecasting accuracy will be a determining factor for whether the revision is acceptable.

Forecasts are needed when outcome labels are unknown. But without outcome labels, accuracy cannot be directly estimated. In practice, forecasting accuracy is evaluated using labeled test data realized from the same joint probability as the training data. A classifier is trained as usual, and test data fitted values are used as forecasts. For each case, the fitted values are compared to the outcome class label to determine accuracy. Key results conventionally are provided in confusion tables. Accuracy is computed for each fitted outcome class as the proportion of cases for which the test data fitted value is the same as the test data outcome label. These are aggregate assessments of accuracy. When real forecasts are made, there are no outcome labels and no way to directly address forecasting accuracy. Accuracy is inferred from the aggregate assessments.

The existing risk assessment tool used by the department of probation and parole had its overall accuracy originally established using forecasting summary statistics a confusion table. Going forward, however, confusion tables and associated statistical summaries can have important limitations addressed later at some length. We will make an important distinction between assessments of forecasting accuracy for a risk tool overall and assessments of forecasting accuracy for individuals.

Conformal prediction (Vovk et al., 2005; Lei et al., 2018; Gupta et al., 2019; Romano et al., 2020; Angelopoulos et al., 2020) is an alternative, generic methodology providing accuracy guarantees for predictions from any machine learning algorithm. In this paper, we emphasize categorical response variables forecasted via nested conformal prediction sets (Gupta et al., 2019). In conformal prediction, each possible set of outcome classes is evaluated to find the “best” prediction set. Unlike forecasting procedures that provide the best single outcome, conformal prediction can return a set of more than one outcome class when a classifier cannot provide sufficiently definitive distinctions between outcome classes. Associated with the best prediction set is an estimate of the probability that the chosen prediction set contains the true outcome. Such claims have finite sample guarantees.

---

\(^1\)We will focus exclusively on probationers who constitute the vast majority of offenders supervised. Parolees are generally supervised at the state level.
One also can construct the best prediction set for particular offender types, defined by their predictor values. Uncertainty claims for these prediction sets have asymptotic guarantees.

Forecasts and associated uncertainties produced by common statistical learning procedures can differ substantially from those produced with the nested conformal approach. Such differences are highlighted in the pages ahead. We argue that nested conformal prediction sets are an instructive, new way to assess forecasting accuracy that, on balance, will often be the preferred approach. We see this paper as a demonstration of concept as plans for a new statistical learning, risk assessment, procedure are formulated. Arguably, nested conformal inference should be a fundamental part of the enterprise.

In Section 2, we describe the policy setting and introduce the data from which forecasts will be made that illustrate the principal statistical issues. In Section 3, we provide some background on conformal prediction and nested conformal prediction for numeric/continuous $Y$. In Section 4, we discuss nested conformal prediction for categorical $Y$, which is of primary interest for our application. We also compare nested conformal prediction to confusion tables and address the discrepancies. In Section 5, we offer an alternative nested conformal method for categorical $Y$ that can corroborate some inferences from a confusion table. Finally, Section 7 offers several broad conclusions and makes recommendations for practice. In short, for criminal justice risk assessments, we consider how one can obtain more informative assessments of forecasting accuracy.

**Notation.** Throughout the paper, we use $1 - \alpha$ to denote the required probability guarantee as in (3) and (4). We also use $\gamma \in [0, 1]$ to denote intermediate probabilities. In our explanations, we use $\delta$ as a value tending to zero and can be thought of as an arbitrary small value. For any event $A$, we write $1_A$ to denote the indicator, i.e., $1_A = 1$ if event $A$ occurs/holds and $1_A = 0$ if event $A$ does not occur.

2. The Policy Setting. For readers unfamiliar with probation supervision provided in urban areas in the United States, we offer some background information, including important details on the local setting. Ahlman and Kurtz (2008) can be consulted for further information. This section helps us set the stage for the data we use as well as the stakeholder concerns/desiderata.

A sentence of probation ordered by a court typically is an alternative to incarceration. A convicted offender is released under supervision, the intensity and content of which can vary widely. For offenders thought to
pose a serious risk to public safety or likely to re-offend, the supervision can be very intrusive: home detention, close surveillance by probation officer, unannounced home or workplace visits, frequent drug tests, GPS monitoring, and more. For offenders seen as low risk, supervision can be in name only. In the extreme, a probationer may only need to report once a month to an electronic kiosk to update a home address, contact phone number, and place of work. However, attached to any sentence of probation can be a variety of conditions such as attending AA meetings, anger management counseling, curfews, community service, restitution, and prohibitions from socializing with certain individuals such as gang members.

Agencies in charge of probation supervision often have wide discretion in how oversight is implemented. Some probation conditions can be relaxed or even removed. New probation conditions can be imposed. Probation officers also have considerable discretion in how strictly the conditions are enforced. For example, a probation officer may choose to overlook one failure to attend an office meeting but not several over 6 months.

Although there can be good reasons for discretion exercised by probation departments and individual probation officers, there are the dual risks of inefficiencies and inconsistencies across probationers. In response to these and other concerns, a large, urban, department of probation and parole began in 2007 using random forests to inform decisions about how to allocate the intensity and content of probation supervision. Their intent was to provide the least restrictive forms of supervision consistent with the public safety and control of crime. A necessary condition was determining the varying needs and risks for each probationer (Ahlman and Kurtz, 2008).

The algorithm’s forecasting accuracy was promising in test data, and approximately 60% of the offenders were projected to be arrest-free while under supervision. But by itself, that was insufficient (Barnes et al., 2010). The implications for agency practices were unclear.

To clarify matters, a randomized field experiment was undertaken to determine whether offenders forecasted to be arrest-free could be supervised at “Low-Intensity” uncompromised by an increase in re-arrests. 1,559 low risk offenders just beginning their probation sentence were identified by the algorithm. Half were randomly assigned to standard supervision and half were randomly assigned to low intensity supervision.

Under the low intensity supervision, case loads for each probation officer were increased on average from approximately 150 offenders to 400 offenders, which automatically meant less frequent and lengthy contacts between probationers and probation officers. For example, the number of office visits per year was reduced on average from 4.5 to 2.4. Drug testing was only to
be done if court ordered. After three failed drug tests, drug treatment was offered as an alternative to incarceration.

All of the offenders in the study were followed for one year through their administrative records. All arrests were recorded. A standard analysis of the experiment showed that the re-arrest proportions were effectively the same for offenders under either supervision regime. For all crimes, 16% of the experimental group was re-arrested compared to 15% of the control group. The usual null hypothesis could not be rejected at the .05 level (p-value = 0.61). Results were much the same for particular crime types such as drug offenses and violent offenses (Barnes et al., 2010). A second analysis using different methods arrived at equivalent results (Berk et al., 2010).

As a result, the department reorganized its practices so that probationers forecasted to be arrest-free were placed in low intensity supervision. Cost savings were allocated to more intensive and expensive supervision for probationers forecasted to be re-arrested. This made the machine learning risk assessments critical. A forecasted outcome class could have dramatic effects on an offender’s experience while on probation and impact the agency’s budget.

Over the past decade, the statistical learning risk instrument was retrained twice with more recent data, but there were no material changes in the overall approach. For example, accuracy was still addressed using confusion tables with test data. During the past year, there has been growing interest in major revisions of the risk instrument to better represent an evolving mix of offenders and governing regulations, and to capitalize on recent technical developments in predictive inference. The former is illustrated by recent increases in shootings and homicides coupled with the volatile politics surrounding police and prosecutorial practices. The latter is a major theme in the pages ahead. Fairness has been addressed at some length in other work (Berk et al., 2020; Berk and Kuchibhotla, 2020). We focus on forecasting accuracy, while making a distinction between the probability that a given forecasted class is correct and the probability that a “best” prediction set is correct.

Description of the Data. Building on our discussion in Section 2, we have data on 102,555 offenders who began a sentence of probation between 2009 and 2013 in the large urban jurisdiction. The substantial number of observations means that asymptotic as well as finite sample guarantees are available. For each offender, there are 2 years of followup data characterizing conduct while on probation. At the urging of probation officials and other stakeholders, we use three outcome variable classes: an arrest for a crime...
of violence, an arrest for a nonviolent crime, and no arrest at all. Concerns about crime typically emphasize crimes of violence. Distinctions between violent crimes and nonviolent crimes are critical because public safety is an essential governmental responsibility.

Seventeen predictors were taken from the usual criminal justice administrative records: rap sheets, charges for the crime(s) that led to the probation sentence, age, gender, and the earliest crime for which the offender was charged as an adult. Race was not included because of objections raised by local stakeholders. Gender was included because no objections were raised. Also, it was widely accepted that men account for the vast majority of violent crimes, such as aggravated assault and homicide, and that commission these offenses is not a product of social disadvantage experienced by men as a group. However, the jurisprudence determining which features of individuals are legally protected is complicated and setting specific. See, for example, the U.S. Civil Rights Act of 1964 (https://www.archives.gov/milestone-documents/civil-rights-act). The final set of predictors were:

- Age: in years
- Gender: male=1, female=0
- Elapsed time since most recent arrest for a violent crime: in years
- Elapsed time since most recent arrest for a weapons crime: in years
- Age at earliest arrest: in years
- Age at earliest arrest for a crime of violence: in years
- Prior jail incarcerations: a count
- Prior prison incarcerations: a count
- Prior murder arrests: a count
- Prior property crime arrests: a count
- Prior violent crime arrests: a count
- Prior drug distribution arrests: a count
- Prior weapons arrests: a count
- Current charges for a violence crime: a count
- Current charges for a drug crime: a count
- Current charges for a property crime: a count
- Current charges for a firearm crime: a count

The data can be seen as realized from a joint probability distribution characterizing all offenders in the relevant jurisdiction sentenced to probation from several years before 2009 to several years after 2013. During that time, there were no major changes in law or administrative procedures that would have materially altered the mix of offenders sentenced or sentencing considerations by local judges. A case for exchangeability is made by noting
that data on each offender was effectively realized independently. Whether one individual was placed on probation was arguably unrelated to whether any other individual was placed on probation. For example, cases are tried one at a time and in this jurisdiction, there are many trial judges presiding in different courtrooms. Independent realizations is sufficient for exchangeability. Datasets $D_1$ ($N = 51278$) and $D_2$ ($N = 51277$) were constructed as random disjoint splits, which leaves exchangeability intact. $D_1$ constitutes training data. $D_2$ constitutes test data.

**Preliminary Data Analysis.** The analysis required asymmetric costs for classification errors. For example, failing to correctly identify an offender who later is re-arrested for a violent crime has very different, and arguably more costly, consequences from failing to correctly identify an offender who is arrest free. The relative cost of different classification errors should be built into a risk algorithm because otherwise, the forecasts will not properly represent the importance of different kinds of errors; the forecasts of risk can be seriously misleading.

Arriving at target cost ratios requires input from stakeholders, and in practice, a consensus about relative costs usually is quickly reached (Berk, 2019). Generally, risk algorithms can be tuned so that target cost ratios are reasonably well approximated in empirical confusion tables. Suppose the target cost ratio for incorrectly classifying an offender, who is actually arrested for violent crime, as not being arrested at all was set at 5 to 1. Likewise, the target cost ratio for incorrectly classifying an offender, who is actually arrested for a nonviolent crime was set at 2 to 1. Finally, suppose the target cost ratio for incorrectly classifying an offender, who is actually arrested for nonviolent crime, as not being arrested at all also was set at 2 to 1. It is important to stress that such ratios represent subjective value preferences that are determined before a risk algorithm is applied. Different groups of stakeholders might well arrive at different cost ratios leading to different risk results. The target cost ratios just specified are reasonable for illustrative purposes.

Algorithmic training was undertaken with the training data $D_1$. Stochastic gradient boosting for a multinomial outcome was applied using the procedure *XGBoost* in R (Chen and Guestrin, 2016). Weighting was introduced to arrive empirically at sufficiently good approximations of the target cost ratios. Table 1 is the resulting confusion table constructed for the $D_2$ data, here serving as the test dataset.

Entries in Table 1 are the case counts. For example, 18,661 subjects out of 51,277 test data subjects are correctly identified by the classifier as
Table 1
Out-of-sample data confusion table (based on $D_2$) for three outcome classes: no arrest, arrest for a nonviolent crime, arrest for a violent crime (N=51,277)

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>No Arrest</th>
<th>Nonviolence</th>
<th>Violence</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>No arrest</td>
<td>18661</td>
<td>8120</td>
<td>3753</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Nonviolence</td>
<td>3617</td>
<td>10274</td>
<td>2410</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Violence</td>
<td>682</td>
<td>1009</td>
<td>2751</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Forecasting Error</td>
<td>0.19</td>
<td>0.47</td>
<td>0.69</td>
<td>0.39</td>
<td></td>
</tr>
</tbody>
</table>

not arrested. Similarly, 8,120 subjects who are not arrested are incorrectly identified by the classifier as arrested for a non-violent crime. Classification error and forecasting error are computed from the test data as follows.

Classification error for a particular outcome, say no arrest, is the proportion of subjects erroneously labeled by the classifier as arrested for a non-violent or violent crime among those who are not arrested. Mathematically,

\[
\text{Classification Error (for no arrest)} := \frac{\sum_{i \in D_2} \mathbb{1}\{\hat{Y}_i \neq Y_i, Y_i = \text{no arrest}\}}{\sum_{i \in D_2} \mathbb{1}\{Y_i = \text{no arrest}\}}.
\]

Here, $Y_i$ is the true outcome for subject $i$ in $D_2$ data, and $\hat{Y}_i$ is the fitted outcome from the trained classifier (i.e., the outcome with the highest estimated probability). The classification error in (1) can be seen as an estimator of $P(\hat{Y} \neq Y|Y = \text{no arrest})$. From Table 1, the classification error for no arrest is estimated to be 0.39.

The forecasting error for a particular outcome, say no arrest, is the proportion of subjects arrested for either a nonviolent or violent crime among the subjects labeled erroneously by the classifier as not arrested. Mathematically,

\[
\text{Forecasting Error (for no arrest)} := \frac{\sum_{i \in D_2} \mathbb{1}\{\hat{Y}_i \neq Y_i, \hat{Y}_i = \text{no arrest}\}}{\sum_{i \in D_2} \mathbb{1}\{\hat{Y}_i = \text{no arrest}\}}.
\]

The notation $Y_i$ and $\hat{Y}_i$ is unchanged. Forecasting error (2) can be seen as an estimator of $P(\hat{Y} \neq Y|\hat{Y} = \text{no arrest})$ from test data realized at the same time as the training data. From Table 1, the estimate for no arrest is 0.19. Uncertainty for that estimate, or any summary statistics from the confusion table, is easily estimated resampling the test data with an X-Y (pairs) bootstrap, here for the subset of cases for which no arrest is forecasted by the classifier. With such a large number of observations, there is very little uncertainty in the estimate of 0.19 (i.e., the standard error is in the third
NESTED CONFORMAL FOR CLASSIFICATION

decimal place). Note however, this is uncertainty for a summary estimate of accuracy. As we emphasize later, this is not the uncertainty for a given forecast which we examine latter with conformal prediction sets.

Classification error is important as a gauge of overall algorithmic performance. Hypothetically, when the true outcome class is known, how often is it incorrectly classified? One can see on the right margin of the table that all three outcome classes were misclassified roughly 40% the time.

When real risk assessments are undertaken, attention shifts to forecasting error. The true outcome class is not known. Rather, a forecast $\hat{Y}$ is computed for each unlabeled case and used to help inform real decisions with real consequences. But training and test data collected earlier for risk instrument development can be used to estimate the chances a highest probability outcome is wrong. Just as for classification error, this is an overall assessment of algorithmic performance, for which standard errors are easily estimated.

Forecasting error, reported at the bottom of the the table, ranged from approximately a fifth of the cases (i.e., 0.19) to about two-thirds of the cases (i.e., 0.69). The latter was substantially inflated by a large number classification errors caused by the high relative costs, specified by stakeholders, of erroneously classifying offenders who were actually violence prone, as posing no threat to public safety. A lower cost ratio would reduce the number of inflated risk estimates, but would substantially increase the number of violence prone offenders being overlooked. There typically are tradeoffs in confusion tables between classification errors and forecasting errors even though they condition on different things.

The likelihood of forecasting errors can be increased when the highest probability forecast is insufficiently definitive. For instance, if for a given offender the estimated probabilities are 0.45 for no arrest, 0.42 for an arrest alleging non-violent crime, and 0.13 for an arrest alleging violent crime, the close proximity of 0.45 and 0.42 can imply that the classifier is unable to make an authoritative distinction between the two outcomes. Yet, for a Bayes classifier, the highest probability class automatically becomes the projected

\footnote{Because the proposed risk instrument revision is motivated in part by changes in the mix of offenders and governing regulations, estimates of classification error and forecasting error from the new data likely would be somewhat different, an issue we will not address in this paper.}

\footnote{Table 1 shows that the stakeholder, target cost ratios were well approximated in the $D_2$ data. For example, the target 5 to 1 ratio when offenders who were re-arrested for a violent crime were classified as not re-arrested had an empirical cost ratio of about 5 to 1 (i.e. $3753/682 \approx 5.5$). In other words, one violence-prone offender misclassified as a risk-free was “worth” approximately five risk-free offenders misclassified as violence prone. It is very difficult to exactly reproduce the target cost ratios because proper confusion tables are constructed from test data, not training data.}
outcome. We show later that with conformal prediction sets, it often will be preferable to forecast more than one outcome class to more properly represent forecasting performance.

Risk assessment results are usefully evaluated in comparison to current practice. The marginal distribution of the outcome classes represents that practice. A sentence of probation is determined by a sitting judge, and all of the offenders in this study were placed on probation; none were incarcerated. No quantitative forecasts of risk were used, but presumably the judge would not mandate probation for convicted offenders thought to be a public safety risk.\(^4\)

Suppose for the moment that implicitly sentencing judges make decisions by the equivalent of a Bayes classifier. According to the empirical marginal distribution of the outcome classes, the “default” forecasts would be no subsequent arrest. In fact, 59% of the offenders are not arrested. But 41% of the offenders were. The 41% can be seen as erroneous forecasts.

With the available data analyzed using stochastic gradient boosting, Table 1 shows that if no arrest is predicted, 19% of the cases would be incorrectly forecasted. This is a meaningful improvement in forecasting accuracy compared to 41%, which can be also characterized as a reduction in forecasting uncertainty. It is bolstered by the cost ratios, making failures to correctly classify new arrests, especially for violent crimes, relatively costly. The boosting algorithm works hard to avoid such errors; the counts of 3617 and 682 are smaller than they would have been if their relative costs were reduced.

There are smaller improvements in accuracy for the other two outcome classes had they been chosen by some implicit forecasting method. If an arrest for a nonviolent crime is forecasted from the marginal distribution of the outcome classes, it would be wrong 68% of the time. The same forecast from the risk algorithm would be wrong 47% of the time. If an arrest for a violent crime is forecasted from the marginal distribution of the outcome classes, it would be wrong 91% of the time. The same forecast from the risk algorithm would be wrong 69% of the time. These are meaningful improvements that reduce forecasting uncertainty in data that might be enhanced substantially with different cost ratio weighting. Still, aggregate forecasting accuracy (i.e., 31% correct) is worse than the 50% accuracy that would be achieved by flipping a fair coin.\(^5\)

\(^4\)A smaller number of offenders were incarcerated, but they are not part of our study because there is no conduct on probation to measure.

\(^5\)With iid data, there are also relatively simple ways to resample the test data such that conventional confidence intervals and statistical tests can used with any of the summary statistics computed from a confusion table. In the interest of space, these will not be discussed here. A readily accessible treatment can be found in Berk (2020, Section 1.4.5).
performance, forecasting error is reduced a nontrivial amount.\footnote{The random, disjoint split into $D_1$ and $D_2$ introduces additional uncertainty into the results. To get sense, we repeated the analysis several times, each with a random reconstruction of $D_1$ and $D_2$. There were noticeable changes in the results that typically improved the aggregate assessments of performance a bit. But, for purposes of this paper, they made no material difference.}

Many stakeholders are more confident about statistical learning forecasts when the dominant predictors comport well with expectations. \textit{XGBoost} provides a measure of each predictor’s “importance.” Importance is measured as the average contribution to the fit for each tree used by the boosting algorithm in mean squared error units, then standardized so that the importance measure over all predictors sums to 100\%. This is an \textit{in-sample} approach distantly related to forecasting performance. There are more appropriate measures in forecasting settings, available with other algorithms \citep{Breiman2001}.

Four predictors of the seventeen are responsible for about half the fit: age, the number of prior jail sentences, the recency of latest prior arrest for a violent crime, and the earliest age at which an offender was arrested. Other analyses using partial dependence plots showed that the risks increase for younger offenders who had a larger number of prior jail sentences, who had an arrest for a violent crime more recently, and who began their criminal activities at an earlier age. No causal claims are implied because an algorithm is not a model, but none of the associations are a surprise \citep{Berk2009, Berk2019}.

Without discussing at some length the criminal justice setting and the policy issues at stake, it is difficult to address whether any imperfect risk algorithm can perform well enough. Much of that discussion depends on what the future holds when the proposal to revise the risk algorithm proceeds. As the Black Lives Matter agenda illustrates, the political environment can be extremely animated. Perceived fairness and transparency will matter as well.

3. Some Background for Conformal Prediction. Statistical learning classifiers commonly are used to make forecasts for single, unlabeled cases. For ease of exposition, we consider for the moment binary categorical response variables. Illustrations include whether a spot on a lung X-ray is a cancer precursor \citep{Yan2018}, whether a mortgage applicant will repay the loan \citep{Chen2020}, or whether a particular hurricane will make landfall \citep{Alemany2019}. One needs procedures that properly determine the precision and uncertainty of such forecasts.

Using the example of a hurricane landfall, the four possible prediction sets are \{Yes\}, \{No\}, \{Yes, No\} and $\emptyset$ (i.e., the empty set). The first could be a very informative if one can claim it is the true outcome class with a
probability such as 0.90. Evacuation orders might follow. The second could be very informative if one can claim it is the true outcome class with a similarly high probability. There would be no evacuation orders. The third is an exercise in statistical humility; the forecasting procedure is unable to choose a single outcome class. This too would be important to know, perhaps making other information more important. The final prediction set can mean that the hurricane is atypical, and that the existing data do not apply. But this too may contain useful information. At a time when the consequences of climate change are rapidly materializing, it would be important to consider whether a particular weather event is unprecedented.

The inferential task in these examples can be formulated in an unconditional manner; one only asks for guarantee on average over all the forecasts. No guarantee is provided for a particular configuration of prediction values.

Returning to a criminal justice setting, consider the task of predicting in a binary fashion whether an individual on probation is arrested for a violent crime based on the predictor vector $X = (X_1, X_2, X_3)$, with $X_1$ as age, $X_2$ as gender, and $X_3$ as the number of prior arrests. A prediction set $\hat{C}(x)$ is said to have an unconditional guarantee if for any future offender with predictor vector $X_f$ and the true response $Y_f \in \{0, 1\}$ one has

$$\Pr \left( Y_f \in \hat{C}(X_f) \right) \geq 1 - \alpha,$$

where $1 - \alpha$ is the specified probability such as 0.90. In words, (3) conveys that if one constructs a $\hat{C}(\cdot)$ for each of 100 future offenders, about $(1 - \alpha)$ of these subjects will have their true response $Y_f$ within their corresponding prediction set $\hat{C}(X_f)$. This is the approach taken by conformal inference, whether the outcome is numeric or categorical, and it has strong finite sample guarantees (Vovk et al., 2005).

Forecasts can also be made in a conditional manner. For example, the distribution for whether there is a re-arrest for a violent crime can differ dramatically between male and female offenders and between younger and older offenders. In contrast to (3), a prediction set $\hat{C}(x)$ is said to have a conditional guarantee if for any future offender with predictor vector $X_f = x$ and the true response $Y_f \in \{0, 1\}$, one has

$$\Pr \left( Y_f \in \hat{C}(X_f) \mid X_f = x \right) \geq 1 - \alpha \quad \text{for all } x,$$

where, again, $1 - \alpha$ is the required probability such as 0.90. In comparison to (3), (4) provides the guarantee for any specific configuration of predictor values, and as before, whatever the forecast happens to be.

---

7 An arrest is coded as “1,” and an absence of an arrest is coded “0.”
To briefly consider the practical importance of the conditional guarantee (4), suppose \( x = (26, \text{Male}, 5) \); the individual is a male of age 26 with 5 prior arrests. Then (4) conveys that if one provides a \( \hat{C}(\cdot) \) for each of 100 future male offenders of age 26 with 5 priors, about \( (1 - \alpha) \) proportion of them will have their true response \( Y \) lie within a particular prediction set \( \hat{C}(26, \text{Male}, 5) \). Although each offender is processed one at a time, there is a distribution on the responses even after fixing the predictors to be \( (26, \text{Male}, 5) \). Nothing specific is offered about a single offender. Still, one has a guarantee for sub-groups of offenders with the same configuration of predictor values.

Although unconditional guarantee (3) can be provided for finite samples as shown in Vovk et al. (2005), Lei et al. (2018), Barber et al. (2019a) proved that the conditional guarantee (4) is in general impossible to attain in finite samples. It is, however, possible to provide confidence sets satisfying the unconditional guarantee (3) in finite samples and to satisfy (4) when the sample size is large enough. These results can be achieved using the nested conformal approach advanced in Gupta et al. (2019, Appendix D); see also Izbicki et al. (2019). This approach is applicable for categorical and numeric responses. There are many settings in which conditional inference is preferable. In our application, it helps to insure that “similarly situated” offenders are treated similarly.

Because of our policy setting, we emphasize nested conformal prediction sets for categorical response variables. Gupta et al. (2019, Proposition 1) prove the formal properties for the nested approach but their discussion focuses largely on numeric outcomes. Although the theory carries over, there are a number of practical considerations for categorical outcomes that need to be unpacked. Because conformal inference is relatively new to many statisticians, we include a substantial didactic material. Readers already familiar with nested conformal inference might wish to skim to Section 4.

**Nested Conformal Prediction Regions for a Numeric \( Y \).** For a numeric/continuous \( Y \), one has prediction intervals. For a categorical \( Y \), one has prediction sets. We build, in particular, on prediction sets that are nested. That is, prediction sets having fewer elements are contained within prediction sets having more elements (e.g., think of the Russian nesting doll). For didactic purposes, however, we begin with a very brief discussion of nested prediction intervals for a numeric \( Y \). In this context, shorter prediction intervals are contained within longer prediction intervals. Details for categorical \( Y \) are given in Section 4.

Suppose one has exchangeable data \( (X_i, Y_i), 1 \leq i \leq n \) from a distribution
and the aim is to construct an interval $\hat{C}(x)$ such that if a new case $(X_{n+1}, Y_{n+1})$ is exchangeable with $(X_i, Y_i), 1 \leq i \leq n, \mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$. Intuitively, smaller prediction intervals for $\hat{C}(x)$ are preferred; they are more precise. But the best prediction interval depends on the “oracle” (i.e., true) conditional density $p(y|X = x) = p(y|x)$ given by

$$\hat{C}_{\text{oracle}}(x) := \left\{ y \mid p(y|x) \geq \tau_{\text{oracle}}(\alpha, x) \right\};$$

where a threshold $\tau_{\text{oracle}}(\alpha, x)$ is chosen to be the largest possible value such that $\mathbb{P}(Y_f \in \hat{C}_{\text{oracle}}(X_f)|X_f = x) \geq 1 - \alpha$. As before, $x$ denotes a set of predictor values used to fit a particular value of $y$. A larger value for $\tau_{\text{oracle}}(\alpha, x)$ translates into a higher threshold that necessarily produces a smaller prediction interval/set, but the best prediction interval/set cannot be computed in practice because the true conditional density $p(y|x)$ and hence, $\tau_{\text{oracle}}(\alpha, x)$, are unknown. Because $p(y|x)$ is unknown and is considered oracle knowledge, we call the prediction set $\hat{C}_{\text{oracle}}(\cdot)$ in (5) as the “oracle” prediction set.

FIG 1. Illustration of the level sets $\{ y : p(y|x) \geq t \}$ for a bimodal conditional density $p(y|x)$ with three illustrative thresholds $t_0, t_1, t_2$.

Estimation is addressed shortly. For now, we illustrate properties of the “oracle” prediction set with a bimodal density. Figure 1 is a plot of the
true density \( p(y|x) \) (for a specific \( x \)) with three hypothetical thresholds at 0, 0.05, 0.10. For the threshold at \( t_0 = 0.0 \) (solid line), the set \( \{ y : p(y|x) \geq t_0 \} = \{ y : p(y|x) \geq 0 \} \) contains all possible \( y \)'s between the two red arrowheads. The coverage probability for this set is 1 because all of the area under the density curve falls above the threshold. Increasing the threshold to \( t_1 \) (in dash-dotted line) excludes some \( y \)-values and concurrently, the coverage guarantee is reduced; the set of \( y \)'s for which the density is above \( t_1 \) is smaller than those with density above \( t_0 \). The same reasoning applies to \( t_2 \) (in dashed line). The set containing the relevant \( y \)'s corresponds to the values between the two brown arrowheads on the left combined with the values between the two brown arrowheads on the right. The set corresponding to the threshold \( t_2 \) is not an interval but rather a union of two intervals.

Increasing the threshold \( t \) to the top of the peak, say \( t = t^* \), leaves no value of \( y \) that satisfies \( p(y|x) \geq t^* \), and hence, \( \{ y : p(y|x) \geq t^* \} \), conceptualized as a set, is empty. The coverage probability for this set is 0. By gradually raising the threshold \( t \) from 0 to \( t^* \) the top of that peak, the coverage probability for \( \{ y : p(y|x) \geq t \} \) drops smoothly from 1 to 0 and equals the required probability of \( 1 - \alpha \) at a particular threshold. That special threshold is \( t_{\text{oracle}}(\alpha, x) \), and is the best prediction set of coverage \( 1 - \alpha \), because it has the smallest volume among all sets that have a coverage of at least \( 1 - \alpha \).

Estimation of the best prediction set is central to all that follows and illustrates fundamental ideas required when shortly we transition to a categorical \( Y \). Consider some estimate \( \hat{p}(y|X = x) = \hat{p}(y|x) \) of the true conditional density \( p(y|x) \). Using this approximation, we obtain an estimate \( \hat{l}(\alpha, x) \) of the oracle threshold \( t_{\text{oracle}}(\alpha, x) \).

The “naive” estimate of the oracle prediction set

\[
\hat{C}_\alpha^{\text{naive}}(x) := \{ y : \hat{p}(y|x) \geq \hat{l}(\alpha, x) \},
\]

cannot have any finite sample guarantee because \( \hat{p}(y|x) \) is an estimate with unknown arbitrary accuracy. But for any \( \gamma \in [0, 1] \), there exists a function \( f(\cdot) \) depending on the estimator \( \hat{p}(\cdot|\cdot) \) such that \( \mathbb{P}(Y_f \in \hat{C}_\gamma^{\text{naive}}(X_f)) \geq 1 - f(\gamma) \).

By making \( \gamma = f^{-1}(\alpha) \),

\[
\mathbb{P}(Y_f \in \hat{C}_\gamma^{\text{naive}}(X_f)) = 1 - f(\gamma) = 1 - f(f^{-1}(\alpha)) = 1 - \alpha.
\]

One has a form of calibration. Because \( \hat{C}_\alpha^{\text{naive}}(\cdot) \) does not have the desired coverage, \( \alpha \) is altered or “calibrated” so that the coverage becomes the original \( 1 - \alpha \). However, \( f(\cdot) \) is unknown and hence, conformal prediction is used to calibrate the coverage. The mathematical details of \( \hat{l}(\alpha, x) \), as well as calibration for a numeric \( Y \), is beyond the scope of this paper. Interested readers should consult Gupta et al. (2019, Appendix D).
With this background and intuition in place, the next section considers
categorical outcome classes in detail.

4. Nested Conformal Prediction Sets for Categorical Y. For a
categorical Y, the basic ideas are similar, but the details differ substantially
because the thresholds are applied to a probability mass function, not a
probability density function, and response classes are on the horizontal
axis. Just as in Figure 1, the threshold can be increased smoothly, but the
probability above the threshold goes from 1 to 0 in abrupt steps. Included
y-classes change abruptly too. For a given value of α, the goal now is to
determine the best prediction set for \( 1 - \alpha \) using calibration. The abrupt
transitions complicate the process.

To be consistent with the probation application, three outcome classes
are used. The reasoning to follow applies equally well when there are two
outcome classes or more than three. However, with two classes, important
ideas are lost, and with more than three classes, foundational ideas can be
obscured by additional details.

Consider first a probability distribution for a categorical Y and a fixed x
with three outcome classes coded as 0, 1, 2. Their probabilities are
\[
p(0|x) = 0.43, \quad p(1|x) = 0.35, \quad \text{and} \quad p(2|x) = 0.22,
\]
displayed in Figure 2 by three vertical lines in gray. Three illustrative thresh-
olds are shown as horizontal lines using the same visual conventions as before.
The vertical axis is in probability units.

For the full oracle prediction set \( \{y \in \{0, 1, 2\} : p(y|x) \geq t^\text{oracle}(\alpha, x)\} \),
where the threshold \( t^\text{oracle}(\alpha, x) \) is chosen so that the region has \( 1 - \alpha \) coverage.
This closely parallels the earlier discussion for a numeric Y. As the threshold
is smoothly increased, the probability above the threshold is the sum of
the class probabilities for the remaining classes. That probability at the red
threshold and above is 0.22 + 0.35 + 0.43 = 1.0. The coverage for \( \{0,1,2\} \) is
1.0. For example, should the specified \( \alpha = 0.30 \), making 0.70 the specified
coverage, that coverage very conservative, and all three outcome classes are
in the prediction set. Clearly, this is an unsatisfying result. The probability
at or above the blue threshold is 0.35 + 0.43 = 0.78. Coverage for \( \{0,1\} \) is
0.78. This prediction set is smaller, but the coverage is still a bit conservative.
Finally, the probability at or above the highest threshold is 0.43. Coverage
for \( \{0\} \) is 0.43, which falls below 0.70, and contravenes the coverage target.

Table 2 provides an arithmetic translation of these relationships. Consider
in column 1 all illustrative thresholds increasing from a value of 0.0 to include
but not exceed a value 0.22. For each, all outcome classes are included.
Fig 2. Illustration of the level sets \( \{ y : p(y|x) \geq t \} \) for a categorical response with three thresholds at 0.22, 0.35, 0.43. Here \( p(y|x) \) represents the conditional probability mass function (pmf) of the response \( Y \) given \( X \) and the response \( Y \) takes value 0 with (conditional) probability 0.43, 1 with probability 0.35, 2 with probability 0.22.
with a coverage of 1.0, which is larger than the threshold value. For all such thresholds, one has a valid but, as before, a very conservative and uninteresting prediction set.

For an arbitrarily small \( \delta \), there are new thresholds beginning at \( \delta \) above 0.22 and ending at, but not exceeding, 0.35. The probability content above any of these thresholds is \( 0.43 + 0.35 = 0.78 \). The included outcome classes are \{0,1\}; \{2\} no longer qualifies. (See Figure 2.)

From thresholds from \( \delta \) above 0.35 up to but not exceeding 0.43, the single class \{0\} remains; the other two classes are eliminated. (See again Figure 2.) Although the remaining probability is still greater than the threshold values, the coverage falls well below the desired 0.70.

For threshold values between 0.43 + \( \delta \) and 1.0, the sum of the class probabilities is equal to 0.0. There are no outcome classes left, and one has the empty prediction set \( \emptyset \). Likely, this result is useless, at least from a policy perspective.

| Threshold \( t \) | \( \{ y : p(y|x) \geq t \} \) | Sum of Class Probabilities |
|-----------------|-----------------|-----------------|
| 0.0             | \{0,1,2\}       | 0.43 + 0.35 + 0.22 = 1.0 |
| 0.10            | \{0,1,2\}       | 0.43 + 0.35 + 0.22 = 1.0 |
| 0.20            | \{0,1,2\}       | 0.43 + 0.35 + 0.22 = 1.0 |
| 0.22            | \{0,1,2\}       | 0.43 + 0.35 + 0.22 = 1.0 |
| 0.22 + \( \delta \) | \{0,1\}       | 0.43 + 0.35 = 0.78 |
| 0.30            | \{0,1\}        | 0.43 + 0.35 = 0.78 |
| 0.35            | \{0,1\}        | 0.43 + 0.35 = 0.78 |
| 0.35 + \( \delta \) | \{0\}       | 0.43 |
| 0.40            | \{0\}        | 0.43 |
| 0.43            | \{0\}        | 0.43 |
| 0.43 + \( \delta \) | \( \emptyset \) | 0.0 |
| 0.5             | \( \emptyset \) | 0.0 |
| 1.0             | \( \emptyset \) | 0.0 |

From this reasoning and, say, \( \alpha = 0.30 \), one can choose any threshold with values between [0, 0.35] and guarantee coverage of at least 0.7. By convention and in the spirit the discussion for a numeric \( Y \), we take the largest threshold satisfying the coverage, and this yields \( t_{\text{oracle}}(0.3, x) = 0.35 \).

**Estimation.** In practice, an estimator for the oracle probability distribution

---

8Think of \( \delta \to 0 \). Here we are talking about the approaching 0.22 from the right.
is required, which means that the class probabilities and thresholds need to be estimated as well. A threshold now is denoted by \( \hat{t}(\gamma, x) \), where \( \gamma \) can be seen as a provisional value of \( \alpha \) such that \( 1 - \gamma \) is a provisional coverage. The term “provisional” is employed because the impact of different \( \gamma \)-values will be examined. Using reasoning much like that for the oracle setting, a given estimator \( \hat{p}(y|x) \) for a classifier for \( y \in \{0, 1, 2\} \), and any \( \gamma \in [0, 1] \), one can obtain \( \hat{t}(\gamma, x) \) and its “naive” prediction set

\[
\hat{C}_\gamma^{\text{naive}}(x) := \{ y \in \{0, 1, 2\} : \hat{p}(y|x) \geq \hat{t}(\gamma, x) \}.
\]

In words, a naive prediction set for a specified \( \gamma \) includes one or more outcome classes, here from \( \{0,1,2\} \), such that their estimated probability is equal to or larger than the provisional \( \hat{t}(\gamma, x) \) for the provisional \( \gamma \).

As noted in Section 3, the naive prediction set does not have a coverage guarantee because \( \hat{p}(y|x) \) is only an estimate. It is possible to do better using calibration. If \( \hat{p}(y|x) = p(y|x) \), then \( \hat{C}_\gamma^{\text{naive}}(x) \) has the \( 1 - \gamma \) coverage guarantee. One implication that a researcher’s choice of classifier can help insofar as the \( p(y|x) \) is better approximated. Moreover, a more accurate estimator \( \hat{p}(\cdot|x) \) can reduce the \( \gamma \) calibration burdens and facilitate a closer representation of the conditional validity (4).

**Calibration using Conformal Prediction.** Rather than trying to alter \( \hat{p}(\cdot|x) \), calibration proceeds directly to goal of the best prediction set, not indirectly by trying to make \( \hat{p}(y|x) = p(y|x) \). The value of \( \gamma \) for \( \hat{C}_\gamma^{\text{naive}}(x) \) is altered such that finite sample \( 1 - \alpha \) coverage is achieved. Stated more formally, calibration seeks a value of \( \gamma \), such that \( \gamma = \gamma(\alpha) \), from which

\[
P\left( Y_f \in \hat{C}_\gamma^{\text{naive}}(X_f) \right) \geq 1 - \alpha.
\]

Because the probability of an event can be approximated by averaging over sample points, \( \gamma(\alpha) \) can be effectively estimated as

\[
\frac{1}{|D_2|} \sum_{i \in D_2} \mathbb{1} \left\{ Y_i \in \hat{C}_\gamma^{\text{naive}}(X_i) \right\} \geq 1 - \alpha.
\]

For an exact finite sample guarantee, the right hand side should be replaced by \( (1 + 1/|D_2|)(1 - \alpha) \); see Lemma 2 of Romano et al. (2019b) for a proof. This small inflation in \( (1 - \alpha) \) is a requirement for valid inference.

**Conformal Scores.** Conformal scores are the building blocks for calibration. As a first approximation, a conformal score conveys how well the point
“conforms” to the data at hand. A more detailed conception is provided later, once its formal foundation is established.

For any \((x, y)\), a conformal score is defined as

\[
s(x, y) := \sup \{ \gamma \in [0, 1] : \ y \in \hat{C}_\gamma(x) \}.
\]

In words, \(s(x, y)\) is the largest \(\gamma \in [0, 1]\) such that the naive prediction set at \(\hat{l}(\gamma, x)\) contains at least one element of \(y\) corresponding to an actual outcome class. For example, the class might be no arrest while on probation. Illustrations are provided shortly in Table 3.

There is nesting because \(\hat{C}_\gamma(x)\) is monotonically decreasing in \(\gamma\), such that for \(\gamma_1 \leq \gamma_2\),

\[
\hat{C}_\gamma(x) \subseteq \hat{C}_{\gamma_1}(x).
\]

Because the naive prediction sets are nested from \(\hat{C}_{\gamma=0}(x) = \{0, 1, 2\}\) to \(\hat{C}_{\gamma=1}(x) = \emptyset\), the threshold can be increased until the largest \(\gamma\) is chosen such that \(y \in \hat{C}_\gamma(x)\) is true. The nestedness of the sets and the conformal score as defined in (6) constitute the nested conformal framework in Gupta et al. (2019).

To help fix the ideas in (6), consider some new illustrative calculations for Table 3 with the goal of computing a proper conformal score. The oracle probability distribution now is treated as a property of a single, realized, exchangeable case, much as in a true forecasting setting. This is fundamentally different from the earlier setting in which Table 2 was placed, although some of the reasoning and calculations will be familiar. Attention centers on the role of \(\hat{l}(\gamma, x)\) for all \(\gamma \in [0, 1]\). The true outcome classes remain \(\{0, 1, 2\}\) and, as before, suppose \(\hat{p}(0|x) = 0.43\), \(\hat{p}(1|x) = 0.35\), and \(\hat{p}(2|x) = 0.22\). These estimates could be the product of a classifier for that single case.

Starting at the top of Table 3, one has a \(\hat{l}(\gamma, x)\) with \(\gamma = 0.0\). In this role, \(\gamma\) is a place-holder for \(\alpha\) determining the provisional prediction set coverage. Moving downward, the last row for which the sum of the three outcome class probabilities at least \(1 - \gamma = 1.0\) determines an initial \(\hat{l}(\gamma, x)\), which applies to all \(\gamma < 0.22\). Because \(\delta\) can be arbitrarily small (\(\delta \to 0\)), \(\hat{l}(0, x) = 0.22\). Similarly, for any \(\gamma \in [0, 0.22)\), \(\hat{l}(\gamma, x) = 0.22\), and the prediction set is the same. Much as for a numeric \(Y\), higher thresholds are preferred, other things equal, because the prediction regions are smaller even though in this case the prediction set does not change for \(\gamma \in [0, 0.22)\).

When \(\gamma = 0.22\), \(1 - \gamma = 0.78\) (i.e. \(0.35 + 0.43\)). By the same reasoning, the last instance when the sum of probabilities in Table 3 is at least 0.78 is for \(\hat{l}(\gamma, x)\) with \(\gamma < 0.57\). One has for \(\gamma \in [0.22, 0.57)\), a different \(\hat{l}(\gamma, x) = 0.35\), and prediction set is \(\{0, 1\}\).
For $\gamma = 0.57$, $1 - \gamma = 0.43$, the last row for which the sum of probabilities in Table 3 is at least 0.43. Hence, $\hat{t}(0.57, x) = 0.43$. A similar argument implies that for $\gamma \in [0.57, 1)$, $\hat{t}(\gamma, x) = 0.43$. Finally, $\hat{t}(1, x) = 1$.

For the calculations summarized in Table 3, one imagines fixing the threshold at the smallest outcome class probability and increasing the value of $\gamma$ until Equation 6 is satisfied. Then, one fixes the threshold at the next highest outcome class probability and again, increases the value of $\gamma$ until Equation 6 is satisfied. This process is repeated until the outcome classes are exhausted.

| $\gamma$ | $t(\gamma, x)$ | $C^\text{naive}_\gamma(x) = \{y : \hat{p}(y|x) \geq t(\gamma, x)\}$ |
|-----------|-----------------|--------------------------------------------------|
| 0.0       | 0.22            | $\{0, 1, 2\}$                                   |
| 0.20      | 0.22            | $\{0, 1, 2\}$                                   |
| $0.22 - \delta$ | 0.22          | $\{0, 1, 2\}$                                   |
| 0.22      | 0.35            | $\{0, 1\}$                                     |
| 0.30      | 0.35            | $\{0, 1\}$                                     |
| $0.57 - \delta$ | 0.35        | $\{0, 1\}$                                     |
| 0.57      | 0.43            | $\{0\}$                                         |
| 0.7       | 0.43            | $\{0\}$                                         |
| $1.0 - \delta$ | 0.43        | $\{0\}$                                         |
| 1.0       | 1.0             | $\emptyset$                                     |

It follows that for Table 3, there are three conformal scores to be computed: $s(x, 0)$, $s(x, 1)$, and $s(x, 2)$. Starting with $s(x, 0)$, one requires the largest $\gamma$ such that 0 belongs in $C^\text{naive}_\gamma(x)$. From Table 3, this is the set of all $\gamma$ such that $0 \in C^\text{naive}_\gamma(x)$ is $[0, 1)$, the supremum of which is 1. Consequently, $s(x, 0) = 1$.

For $s(x, 1)$, the set of all $\gamma$ such that 1 belongs in $C^\text{naive}_\gamma(x)$ is $[0, 0.57)$, the supremum of which is 0.57. Consequently, $s(x, 1) = 0.57$.

Finally, the set of all $\gamma$ such that 2 belongs in $C^\text{naive}_\gamma(x)$ is $[0, 0.22)$, the supremum of which is 0.22. Consequently, $s(x, 2) = 0.22$.

In short, the three conformal scores, one for each outcome class from $\{0, 1, 2\}$, are

$$s(x, 0) = 1, \quad s(x, 1) = 0.57, \quad \text{and} \quad s(x, 2) = 0.22.$$  

Such reasoning conveys why $s(x, y)$ is called a conformity score. For classifiers providing output probabilities, if $\hat{y}$ is the chosen label for some $x$ (i.e. the fitted class is the same as the observed class), $s(x, \hat{y})$ will have the largest value among the outcome class conformal scores. Here, from the estimated probability distribution, 0 is the most likely outcome and hence, $(x, 0)$
conforms most closely to a classifier’s highest probability selection. It is appropriate, therefore, that \( s(x, 0) \) is the largest conformal score. The second most probable \( s(x, 1) \) conforms less well, and the third most probable \( s(x, 2) \) conforms least well.

More formally, let \( \hat{\pi}(0), \hat{\pi}(1), \hat{\pi}(2) \in \{0, 1, 2\} \) re-define the class labels such that

\[
\hat{p}(\hat{\pi}(0)|x) > \hat{p}(\hat{\pi}(1)|x) > \hat{p}(\hat{\pi}(2)|x).
\]

The new outcome classes are arranged in order from the largest estimated probability to the smallest and re-labeled respectively as \( \hat{\pi}(0), \hat{\pi}(1), \hat{\pi}(2) \). For example, an arrest for a nonviolent crime has the second largest probability, it is denoted by \( \hat{p}(\hat{\pi}(1)) \). Then,

\[
\begin{align*}
s(x, \hat{\pi}(0)) &= 1, \\
s(x, \hat{\pi}(1)) &= \hat{p}(\hat{\pi}(1)|x) + \hat{p}(\hat{\pi}(2)|x), \\
s(x, \hat{\pi}(2)) &= \hat{p}(\hat{\pi}(2)|x).
\end{align*}
\]

For a concrete understanding of the formulae (7) and (8), consider an example case where the estimated probabilities of outcome classes are

\[
\hat{p}(0|x) = 0.27, \quad \hat{p}(1|x) = 0.54, \quad \text{and} \quad \hat{p}(2|x) = 0.19.
\]

For this study subject, the class with largest probability is 1, the class with second largest probability is 0, and the class with third largest (or the smallest) probability is 2. This means \( \hat{\pi}(0) = 1, \hat{\pi}(1) = 0, \) and \( \hat{\pi}(2) = 2 \) in (7). Applying (8) yields

\[
\begin{align*}
s(x, \hat{\pi}(0)) &= s(x, 1) = 1, \\
s(x, \hat{\pi}(1)) &= s(x, 0) = 0.27 + 0.19 = 0.46, \\
s(x, \hat{\pi}(2)) &= s(x, 2) = 0.19.
\end{align*}
\]

Note that computing the scores does not require knowledge of the observed outcome. In this example, if the observed outcome is 0, then the conformal score for the observation \( (x, 0) \) is \( s(x, 0) = 0.46 \).

If any two of the probabilities in (7) are equal, the formula for \( s(x, y) \) is slightly complicated, but one can always add a small amount of noise to all the probabilities so that no two probabilities are equal. This can be done without changing the predictions or prediction sets.
Calibration. With one conformal score in hand for each $D_2$ case, calibration is undertaken by calculating an appropriate quantile. For all $i \in D_2$, and conformal scores $s(X_i, Y_i)$, $\hat{\gamma}(\alpha)$ is defined by

$$1 - \hat{\gamma}(\alpha) := \left(1 + \frac{1}{|D_2|}\right) (1 - \alpha)\text{th quantile of } 1 - s(X_i, Y_i), i \in D_2.$$ 

The best prediction set is

$$\hat{C}_\alpha(x) := \{y : s(x, y) \geq \hat{\gamma}(\alpha)\} = \left\{y : \hat{p}(y|x) \geq \max_{\gamma < \hat{\gamma}(\alpha)} \hat{f}(\gamma, x)\right\}.$$

Because $s(x, y)$ is a conformity score indicating how well $(x, y)$ conforms to the $D_2$ classifications, $1 - s(x, y)$ actually is a non-conformity score. The best prediction set $\hat{C}_\alpha(x)$ is the collection of all $y$ that conforms the best to the actual $D_2$ outcome classes within the required coverage. This is accomplished by taking only those $y$ for which $1 - s(x, y)$, the non-conformity score, is less than $1 - \hat{\gamma}(\alpha)$.

The earlier discussion and the nested conformal method for categorical $Y$ is summarized by Algorithm 1. It provides a scaffold for coding as well as a short summary for the many details examined earlier in service of computing prediction sets used to evaluate a risk classifier’s accuracy. As such, one has an operational “take-home” message formally justified by Theorem 1 immediately below.

**Algorithm 1:** Nested conformal prediction for classification

**Input:** Data splits $D_1$ and $D_2$, coverage probability $1 - \alpha$.

**Output:** A prediction set $\hat{C}_\alpha(\cdot)$ such that $P(Y_i \in \hat{C}(X_i)) \geq 1 - \alpha$.

1. Train a classifier $\hat{p}(\cdot|\cdot)$ on $D_1$. This gives a probability distribution (estimator) for the outcomes from each $x$.
2. Calculate the conformal scores $s(X_i, Y_i)$ for all $i \in D_2$ defined in (6).
3. Compute the $(1 + 1/|D_2|)(1 - \alpha)$-th quantile of $1 - s(X_i, Y_i), i \in D_2$. Call this quantile $1 - \hat{\gamma}(\alpha)$. This is essentially the $[(|D_2| + 1)(1 - \alpha)]$-th largest value in the sequence $1 - s(X_i, Y_i), i \in D_2$
4. return the prediction set

$$\hat{C}_\alpha(x) := \{y : s(x, y) \geq \hat{\gamma}(\alpha)\}.$$

**Theorem 1.** If the data $(X_i, Y_i), i \in D_2$ are exchangeable, then the prediction set $\hat{C}_\alpha(\cdot)$ obtained from Algorithm 1 satisfies

$$P\left(Y_i \in \hat{C}_\alpha(X_i)\right) \geq 1 - \alpha.$$
Proof. The result follows from Proposition 1 of Gupta et al. (2019).

Remark 4.1 The approach taken in Algorithm 1 is called the split conformal prediction method (Papadopoulos et al., 2002; Lei and Wasserman, 2014). There are several, somewhat more involved, versions of conformal methods that do not require sample splitting, and can in principle make better use of the data: jackknife+, CV+, subsampling, and bootstrap. Even a brief discussion of these procedures would take us far afield, and excellent treatments are easily found in Barber et al. (2019b), Gupta et al. (2019), and Kim et al. (2020).

Applying Algorithm 1 to Probation the Data. Tables 4 and 5 provide illustrations of the key features of Algorithm 1. Later, connections to the classifier output in Table 1 will be addressed.

Table 4
Examples of conformal scores based on the estimated class probabilities and the actual outcome class (observed $Y$) for six randomly chosen test subjects from $D_2$. The table shows $s(X_i, Y_i)$ for six random $i$’s in $D_2$.

| Observed $Y$ | $\hat{p}$(NoArrest|x) | $\hat{p}$(NonViolent|x) | $\hat{p}$(Violent|x) | Conformal Score |
|-------------|----------------------|------------------------|---------------------|----------------|
| NoArrest    | 0.58                 | 0.25                   | 0.17                | 1.0            |
| NonViolent | 0.49                 | 0.32                   | 0.18                | 0.50           |
| NoArrest    | 0.47                 | 0.35                   | 0.18                | 1.0            |
| Violent    | 0.18                 | 0.26                   | 0.58                | 1.0            |
| NonViolent | 0.41                 | 0.37                   | 0.21                | 0.59           |
| NoArrest    | 0.27                 | 0.54                   | 0.19                | 0.46           |

From Step 2 in the algorithm, Table 4 shows, for a random subset of six $D_2$ cases, the construction of conformal scores as described in Equation (8). Each row of Table 4 represent a single test subject from $D_2$. The far left column contains the observed outcome class: “NoArrest” for no arrest, “NonViolent” for an arrest for a nonviolent crime, and “Violent” for an arrest for a violent crime. The three columns to the right contain the estimated outcome class probabilities case by case. Conformal scores are provided on the far right. For example, the offender shown at the bottom of the table had no re-arrest while on probation, yet an arrest for a nonviolent crime was forecasted by the classifier because that outcome class had the largest estimated probability. Therefore, the conformal score becomes $0.27 + 0.19 = 0.46$; see Equation (9) and the following discussion.

Figure 3 is a histogram of the conformal scores $s(X_i, Y_i), i \in D_2$, transformed into non-conformal scores $1 - s(X_i, Y_i), i \in D_2$. (Recall that $D_2$ has 51277 subjects.) There are three distinct of groups of scores. On the far
left are the scores for cases in which the class with the highest probability corresponds to the observed outcome class. These are the cases that conform the most closely to the trained classifier and hence, have a non-conformity score of 0. The scores in the middle are for cases in which the observed outcome class has the second largest fitted probability. The scores to the far right are for cases in which the observed outcome class had the smallest fitted probability. The degree of conformity declines from left to right. The large spike at zero is consistent with the earlier results from the confusion table from the stochastic gradient boosting classifier and are also expected because the classifier should do better than chance if there is a signal in the data.

Algorithm steps 3 and 4 are illustrated in Table 5 using a random set of four $D_2$ cases as if the true outcome were unknown. With $\alpha = .05$, a high probability coverage of 0.95 is being sought. The $\tilde{\gamma}(\alpha)$ of 0.26 is relatively small making it more likely that a larger number of outcome classes are included in a prediction set. The column on the far right shows two prediction sets with two elements and two prediction sets with three elements. For $\alpha = .30$ to the immediate left, a moderate coverage probability of 0.70 is preferred. The $\tilde{\gamma}(\alpha) = 0.60$, which is a more demanding criterion for precision. There are two prediction sets with two elements and two prediction sets with a single element. A larger value of $\alpha$ favors smaller prediction sets with
Nested conformal prediction sets for four examples when \( \alpha = 0.30 \) and \( \alpha = 0.05 \). We use the designation \( 0 = \text{No Arrest} \), \( 1 = \text{NonViolent Arrest} \), and \( 2 = \text{Violent Arrest} \). The first column is the probabilities from the classifier, the second column is the conformal scores computed based on (7). The third and fourth columns show the resulting prediction sets from Algorithm 1 when \( \alpha = 0.30 \) and when \( \alpha = 0.05 \). When \( \alpha = 0.30 \), the calibrated \( \hat{\gamma}(\alpha) \) in step 3 of Algorithm 1 is 0.6 and when \( \alpha = 0.05 \), \( \hat{\gamma}(\alpha) = 0.26 \).

| Probabilities \( \hat{p}(0|x), \hat{p}(1|x), \hat{p}(2|x) \) | Conformal Scores \( s(x,0), s(x,1), s(x,2) \) | Prediction set \( \alpha = 0.3, \hat{\gamma}(\alpha) = 0.6 \) | Prediction set \( \alpha = 0.05, \hat{\gamma}(\alpha) = 0.26 \) |
|---|---|---|---|
| (0.34, 0.27, 0.38) | (0.61, 0.27, 1.00) | [2, 0] | [2, 0, 1] |
| (0.19, 0.24, 0.56) | (0.19, 0.44, 1.00) | [2] | [2, 1] |
| (0.60, 0.24, 0.15) | (1.00, 0.40, 0.15) | [0] | [0, 1] |
| (0.30, 0.35, 0.34) | (0.30, 1.00, 0.65) | [1, 2] | [1, 2, 0] |

smaller probability guarantees. One has greater precision but less certainty.

Because the value of \( \alpha \) is chosen by the data analyst, it offers some control over the properties of the forecasts. Should prediction sets with fewer elements be preferred at the cost of smaller coverage probabilities? For criminal justice applications such as ours, the answer will come from stakeholders. For example, stakeholders might be especially concerned about offenders who are forecasted to commit violent crimes. Nonviolent crimes might be viewed primarily as a public nuisance, undesirable to be sure, but not worth increasing incarceration rates. A larger value of \( \alpha \) might then follow. Yet, some stakeholders may be uncomfortable with greater uncertainty. For them, a smaller value of \( \alpha \) might be preferred even if the prediction sets are larger. Such tradeoffs must be addressed and resolved before a risk algorithm is deployed.

These issues apply equally to forecasts with a conditional guarantee (4). In practice, conditional forecasts can be desirable because probation decisions are made one individual at a time. For instance, if re-arrests differ on the average between male and female offenders, conditional forecasts offer a more responsive, and arguably more fair, approach. But the coverage guarantees now are only asymptotic.

**Comparison of Conformal Prediction and Confusion Tables** In previous sections, we have presented two different methods of understanding forecasting accuracy. In Table 1, each case was assigned the single outcome label with the largest probability computed by the classifier. Forecasts for individual offenders then were easily determined, and because the true outcome classes in the test data were known, aggregate forecasting accuracy for each outcome class was easily estimated. But, no distinctions were made between decisive classifications and classifications that are little different...
from a flip of a fair coin. In this setting, conformal inference conveys several advantages.

First, conformal prediction sets do not force a single forecasted class when that forecast is insufficiently reliable; the forecasts are properly cautious. A message of “can’t tell” can impart a very useful result that, for instance, can motivate decision makers to seek additional sources of information before ruling or weight other criminal justice priorities more heavily.\footnote{Commonly, there can be several weeks or more between a conviction or guilty plea and sentencing. There is time to gather more information.} For example, if there are concerns about “mass incarceration,” imprisonment might be reserved solely for the minority of cases in which there is a single forecasted class projecting danger to the public. Such alternatives also might counter rhetorical claims that the risk algorithm is little different from a lottery.

Second, conformal prediction sets can be computed for different subgroups defined by their predictor values. Stakeholders often note that there is considerable heterogeneity in the backgrounds of convicted offenders. One size should not be permitted fit all and indeed, can be an easily identified example of unfairness.

Third, the selected coverage probability can be used as a tuning parameter that trades precision (i.e., the number of elements in the prediction set) against certainty. The preferred tradeoff can be legitimately varied depending on the setting and predispositions of decision-makers and stakeholders. In the end, useful control can be exercised over the forecasting procedures.

Fourth, the nested unconditional prediction sets have valid coverage in finite samples. They also are agnostic to the choice of classifier. Such flexibility can broaden the range of appropriate applications. Conditional forecasts are only justified asymptotically, but at least in criminal justice applications in metropolitan areas, the number of available observations is likely to be very large.

Still, some may argue that when decisions need to be made about individual offenders, allowing for prediction sets with more than one element is a weakness of conformal prediction. Decision makers may become confused, and the credibility of algorithmic risk forecasts may be undermined. Also, too much uninformed discretion can remain in settings where life changing decisions are often made in a matter of minutes.

For these critics, it arguably is better to rely exclusively on confusion tables constructed from test data that provide estimates of forecasting accuracy for subgroups defined by their forecasted class. Beyond forecasts for individuals that seem easier to understand, forecasting accuracy is allowed to differ depending on that forecasted class. Because there can different
concerns about the accuracy of different forecasts, circumstances can arise in which conditioning on the forecast is helpful. Returning to the earlier discussion of the relative costs of forecast errors, certain stakeholders might be especially concerned about forecasting errors for violent crime arrests. In short, conformal inference skeptics might be prepared to use all classifier forecasts, even those that represent equivocal classifier performance.

5. Localized Conformal Prediction. There is another option. We offer a possible compromise between single forecasted labels from confusion tables and conformal prediction. Building on the conventions used by classifiers, the test data can be separated into subsets, each determined by the outcome class with the largest probability. Such subsets are the same as those in each column of a conventional confusion table, but the nested conformal procedure is applied separately to each subset. For the probation data, one would apply the nested conformal procedure separately three times: on the subset of test data determined by the classifier’s “NoArrest” forecast, on the subset of test data determined the classifier’s “NonViolent” forecast, and on the subset of test determined by the classifier’s “Violent” forecast. One will obtain prediction sets as usual, but now separately for subsets of cases defined by a classifier's highest probability forecast.

We call this method “localized conformal prediction,” summarized in Figure 4 for the probation data. Algorithm 1 is applied three times, once to each test data subset, to produce three thresholds $\gamma_0(\alpha)$, $\gamma_1(\alpha)$, and $\gamma_2(\alpha)$.

![Figure 4. Illustration of Localized Conformal Prediction With Each Test Data Subset Determined by the Classifier’s Forecast](image-url)
The three potential prediction sets denoted by $j$ are,

$$
\hat{C}_{j,\alpha}(x) = \{y \in \{0, 1, 2\} : s(x, y) \geq \gamma_j(\alpha)\}, \quad \text{for} \quad j \in \{0, 1, 2\}.
$$

For any future subject with the covariate vector $x$, we first use the classifier from the training data to obtain the forecast $\hat{Y}$. If $\hat{Y} = 0$, one should report the prediction set $\hat{C}_{0,\alpha}(x)$; if $\hat{Y} = 1$, one should report the prediction set $\hat{C}_{1,\alpha}(x)$; if $\hat{Y} = 2$, one should report the prediction set $\hat{C}_{2,\alpha}(x)$. Succinctly, the appropriate prediction set is $\hat{C}_{\hat{Y},\alpha}(x)$.

The prediction set constructed in this manner is more selective than the output of Algorithm 1 applied on the complete test data $D_2$. Nevertheless, the prediction set $\hat{C}_{\hat{Y},\alpha}(x)$ retains the major asset of a finite sample, unconditional guarantee (3), irrespective of the accuracy of the classifier $\tilde{p}(\cdot | \cdot)$.

**Theorem 2.** If the data $(X_i, Y_i), i \in D_2$ are exchangeable, then the prediction set $\hat{C}_{\hat{Y},\alpha}(\cdot)$ obtained consistent with Figure 4 satisfies

$$
P\left(Y_f \in \hat{C}_{\hat{Y},\alpha}(X_f)\right) \geq 1 - \alpha.
$$

This result follows from the arguments of Lemma 1 of Romano et al. (2019a); also, see Section 4 of Vovk (2012). The localized conformal prediction algorithm is one of the many variants of Algorithm 1 leading to finite sample prediction sets that can be more responsive to particular applications. Such variants can be found in Hechtlinger et al. (2018), Sadinle et al. (2019), and Guan and Tibshirani (2019).

Note, however, that the initial subsetting of the data determined by the forecasted class label still makes no distinction before decisive classifications and equivocal classifications. Especially when classification performance is weak, the calibration conformal score distribution that applies to a given case often will be selected by the rough equivalent of a coin flip; the reference distribution could easily have come from a different subset of the data. And because these calibration distributions can differ dramatically over data subsets, their computed prediction sets can vary dramatically too.

**Application of Localized Conformal Method to Probation Data.** To illustrate localized conformal method, we consider its performance compared to the conventional confusion table approach. The reference confusion table is Table 1.

---

[10]Here again we use the designation: “0 = NoArrest”, “1 = NonViolent”, and “2 = Violent”.
Working with a usual confusion table, suppose a data analyst constructs prediction sets containing solely the highest probability forecast. This follows from the test data subsetting described above. Forecasting errors (2) provide the miscoverage probabilities conditional on the forecast. For instance, from the last row of Table 1, when the forecast is “NoArrest,” the true outcome is no arrest with an estimated probability of $1 - 0.19 = 0.81$. Consequently, reporting a single element prediction set would lead to an 81% certainty when the true forecast is “NoArrest.” But for the cases with a forecast of non-violent arrest, the single element prediction set has a certainty of $1 - 0.47 = 0.53$. Single element prediction sets can have different coverages across different forecasted classes, which responds to the concerns just raised; the accuracy of some forecasts matters more than others. In short, confusion tables fix the number of elements in the prediction set to one and commonly lead to different coverage levels for different forecasted outcome classes.

Conformal prediction methods, in contrast, fix coverage at a desired level of $1 - \alpha$ and find a set of one or more outcome classes such that coverage is at least $1 - \alpha$. The nested conformal prediction method (described in Algorithm 1) does not distinguish between cases depending on their forecast. The localized conformal prediction method (illustrated in Figure 4) distinguishes between cases depending on their forecast, similar to the confusion table. This is the key distinction between these two conformal methods.

Both the use of conformal prediction sets or inferences from confusion tables can be formally justified. In practice, the data analyst needs to decide which bundle of forecasting attributes is more instructive for the decisions to be made. For example, how important is it to incorporate information on a forecasted class’s reliability compared to showing the ways in which different forecasted classes can have different forecasting accuracy? If the implications of reliability dominate, nested prediction sets might be preferred. If the implications of differential forecasting accuracy dominate, confusion table approaches might be preferred.

Localized conformal inference offers a compromise. With forecasting accuracy characterized by the number of elements in a prediction set at a predetermined coverage probability, one can compare the distributions of predicted sets across each of the classifier’s highest probability outcome classes. Distributions with their mass over test data concentrated at prediction sets with fewer elements can be seen as more accurate.

A related question is the relative forecasting performance of nested prediction sets compared to localized conformal prediction sets. To get an initial sense of the issues, in Table 6, we compare the nested conformal to localized conformal prediction sets for the four example cases shown in Table 5.
Table 6
Comparison of nested conformal and localized conformal prediction sets for the four examples in Table 5 when $\alpha = 0.30$ and $\alpha = 0.05$. We use the designation 0 = No Arrest, 1 = NonViolent Arrest, and 2 = Violent Arrest. The first and third columns show the nested conformal prediction sets from Algorithm 1 when $\alpha = 0.30$ and when $\alpha = 0.05$; these are repeated from Table 5. The second and fourth columns show the localized conformal prediction sets from Figure 4 when $\alpha = 0.30$ and $\alpha = 0.05$, respectively. The calibrated quantile $\gamma(\alpha)$ for nested conformal are mentioned in Table 5. There are three calibrated quantiles for each $\alpha$ for localized conformal. For $\alpha = 0.3$, $\gamma_0(\alpha) = 0.99, \gamma_1(\alpha) = 0.58, \gamma_2(\alpha) = 0.32$. For $\alpha = 0.05$, $\gamma_0(\alpha) = 0.45, \gamma_1(\alpha) = 0.26, \gamma_2(\alpha) = 0.24$.

<table>
<thead>
<tr>
<th>Prediction set when $\alpha = 0.3$</th>
<th>Localized Prediction set when $\alpha = 0.3$</th>
<th>Prediction set when $\alpha = 0.05$</th>
<th>Localized Prediction set when $\alpha = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{2, 0}</td>
<td>{2, 0}</td>
<td>{2, 0, 1}</td>
<td>{2, 0, 1}</td>
</tr>
<tr>
<td>{2}</td>
<td>{2, 1}</td>
<td>{2, 1}</td>
<td>{2, 1}</td>
</tr>
<tr>
<td>{0}</td>
<td>{0}</td>
<td>{0, 1}</td>
<td>{0}</td>
</tr>
<tr>
<td>{1, 2}</td>
<td>{1, 2}</td>
<td>{1, 2, 0}</td>
<td>{1, 2, 0}</td>
</tr>
</tbody>
</table>

lower level of confidence ($0.7 = 1 - 0.3$), conformal and localized prediction sets are no different, for the four examples presented. At a higher level of confidence ($0.95 = 1 - 0.05$), localized prediction sets are smaller than conformal prediction sets, when the forecast is “NoArrest” (or 0). Precision differs in this instance, and in the probation policy setting, a prediction set with a single element can substantially simplify subsequent decisions. Informally, the localized conformal method accounts better for the heterogeneity in forecasting precision and can yield greater precision than the nested conformal method for some outcome classes.

A more extensive comparison is provided in Table 7. In this table, we report the proportion of cases (for each forecasted outcome) when the conformal and localized methods lead to a prediction set with one element, two elements, or three elements. For instance, the first row of Table 7 shows the proportion of test subjects for whom the highest probability forecast is 0 (= “NoArrest”). So, 0.835 is the proportion of subjects for whom the nested conformal method returned a single element prediction set among the subjects for whom the highest probability forecast is 0 (= “NoArrest”). Prediction sets with a single member are desirable, especially in policy settings when decisions need to be made.

We make the following observations from Table 7. First, there is the tradeoff between $1 - \alpha$ and precision. Comparing the top panel to the bottom panel, prediction sets with a single element are far more common in the top panel for which $1 - \alpha = 0.70$. Prediction sets with three elements are far more common in the bottom panel for which $1 - \alpha = 0.95$. These patterns, which
Table 7
Proportions of obtained prediction sets of particular sizes for nested and localized conformal method when \( \alpha = 0.3 \) and \( \alpha = 0.05 \). For both nested and localized conformal methods, we computed the prediction sets for 51,277 test data points when \( \alpha = 0.3 \) and \( \alpha = 0.05 \). When \( \hat{Y} = 0 \), both methods can only report prediction sets \( \{0\}, \{0, 1\}, \{0, 2\} \) and \( \{0, 1, 2\} \). Similarly, for \( \hat{Y} = 1 \), the possible prediction sets are \( \{1\}, \{1, 0\}, \{1, 2\} \) and \( \{1, 0, 2\} \). For each forecast and each method, we report the proportion of cases for which the obtained prediction set has one element (denoted by “One”), two elements (denoted by “Two”), and three elements (denoted by “Three”). Singletons (sets with one outcome) are the most informative, two element set are the next most informative, and finally, three element sets are the least informative. The proportions are rounded-off to three digits for clarity.

<table>
<thead>
<tr>
<th>( \hat{Y} )</th>
<th>Nested Conformal ((\alpha = 0.3))</th>
<th>Localized Conformal ((\alpha = 0.3))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td>0</td>
<td>0.835</td>
<td>0.165</td>
</tr>
<tr>
<td>1</td>
<td>0.720</td>
<td>0.280</td>
</tr>
<tr>
<td>2</td>
<td>0.624</td>
<td>0.376</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \hat{Y} )</th>
<th>Nested Conformal ((\alpha = 0.05))</th>
<th>Localized Conformal ((\alpha = 0.05))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.798</td>
</tr>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.627</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.510</td>
</tr>
</tbody>
</table>

dominate Table 7, make it difficult to isolate the impact of the conformal method (i.e., the columns). But the impact of the \( 1 - \alpha \) chosen is consistent with the tradeoffs in conformal inference methods.

Within either the top or bottom panel, there are not for this analysis consistent differences in precision between the two conformal methods. Sometimes the nested conformal approach is more precise, and sometimes the localized conformal approach is more precise. For forecasts of no re-arrest, localized conformal inference performs better. For example, in the top panel, the localized method always arrives at prediction sets with a single element. But for the less commonly observed outcomes (i.e., “Violent” and “NoViolent”), the nested conformal approach arguably is superior, particularly for a violent crime arrest, the least common outcome classes.

These differences in precision can be linked to the difference in the guarantees provided by the nested and localized conformal methods. Recall from Figure 4 that the localized conformal approach applies the nested conformal algorithm separately at the same \( 1 - \alpha \) level to three subsets of test data, each corresponding to the highest probability outcome forecasted. Concomitantly, the localized conformal approach provides a coverage guarantee of at least \( 1 - \alpha \) for each of these subsets, despite far smaller numbers of observations for
the two arrest outcome classes from which prediction sets are derived. It is possible to apply the nested algorithms at different levels $1 - \alpha_0$, $1 - \alpha_1$, and $1 - \alpha_2$ for each subsets. The choice of $\alpha_0$, $\alpha_1$, and $\alpha_2$ is mostly a stakeholder’s decision.

Furthermore, the stakeholder cost ratios in this analysis introduce more error into forecasts for both of the two arrest outcomes, especially for violent crime arrests. The cost ratios incentivize the risk algorithm to especially avoid misclassifying offenders who are at high risk for re-arrest. One result is that for these outcome classes, there are many more cases incorrectly labeled as violent, which introduce greater forecasting error. The localized conformal nested approach is undermined by these frequent misclassifications when working with one subset of the data at a time. The nested approach that uses all the data at once is not disadvantaged in this manner. The difference of guaranteeing coverage on each subset versus on all of test data makes it harder to directly compare localized and nested conformal methods; with a stronger guarantee of localized conformal method comes a larger prediction set.

6. Forecasting in Practice with Nested Conformal Prediction Sets. In risk assessment applications, once a classifier has been trained and approved, forecasting is the primary goal. With the results of the trained classifier fixed, test data are used to construct the reference distribution for the nested conformal scores. A value for $\alpha$ is determined by the data analyst. Pseudocode is provided by Algorithm 1.

For each new case, fitted probabilities are computed using the trained classifier. The highest probability outcomes class is determined. Nested conformal scores are then computed for each possible outcome class. These are mapped to a reference, nested conformal score distribution constructed earlier from test data. If a given score falls inside of the prediction region, its outcome class becomes of member of the prediction set. It is a true outcome with $1 - \alpha$ coverage, the value of $\alpha$ previously determined by the data analyst. This is consistent with inference for conventional conformal prediction sets. All that differs is how conformal scores are computed.

In this manner, the best prediction set is assembled. It has finite sample guarantees when the forecasting is unconditional, and asymptotic guarantees when the forecasting is conditional. Localized nested conformal prediction sets can be computed in the same manner, but separately for each data subset defined by a classifier’s highest probability forecast. Procedures are available in R from which a script for forecasting with nested conformal prediction sets can easily be written. But, we know of no existing library in
7. Conclusions. There is an important distinction between forecasting accuracy as part of the overall performance of a risk algorithm and forecasting accuracy for individual cases. Aggregate performance evaluations constructed from test data confusion tables can make a credible case correctly that a particular algorithmic risk assessment is more accurate than the means by which probationer risk levels had previously been determined. If supervision intensity were substantially informed by the algorithmic forecasts instead of current practice, many forecasting errors could be eliminated. Those errors have consequences for supervising offenders with the least restrictive means consistent with public safety and crime control. But this is old news addressed in earlier studies (Barnes et al., 2010; Berk et al., 2010).

The new news is that conformal prediction in general, and nested conformal prediction sets in particular, has the capacity to improve aggregate assessments, but also can be seen as an advance for assessing forecasting accuracy for individuals. Perhaps most important, they can contravene lottery-like forecasts with legitimate “can’t tell” outcomes. Moreover, one has best prediction sets that address both precision and coverage. For unconditional prediction sets, one has valid finite sample coverage. For conditional prediction sets, one has valid asymptotic coverage. If as a policy matter, forecasts are needed for data subsets defined by a highest probability outcome class, localized conformal prediction sets can be constructed that condition on a classifier’s forecasted outcome class. For individuals, however, important lottery-like problems can remain.

In practice, all algorithmic risk results are only as good as the exchangeable data available and the forecasting skill of the algorithm. Criminal justice administrative data that are routinely accessible will often suffice, although their quality and depth can vary by jurisdiction and the criminal justice decision to be made. But whatever the data may be, credible arguments must be made to justify exchangeability.

For the jurisdiction whose data were used in this paper, it remains to be seen going forward whether conformal prediction methods can be accepted by criminal justice decision makers and stakeholders. So far, key officials in the relevant city department strongly favor proceeding. Refitting the classifier with current data would surely be a major improvement. Greater accuracy and more emphasis on communicating uncertainty also would be very useful. However, there is considerable resistance in general to algorithmic risk forecasting as currently practiced or proposed, fueled significantly by misunderstandings about the procedures. Fairness and transparency matter.
too. These complications are locally salient and will shape prospects for any revisions of the probation risk methods. At the very least, there is important educational work ahead.

References.


