Type I and Type II Error Probabilities in the Courtroom

Shin Kanaya* and Luke Taylor†

May 2023 (first version: May 2020)

Abstract

We estimate the probability of a miscarriage of justice using a novel nonparametric estimator for misclassified binary choice models. We show how to test the required large support condition, investigate the validity of the exclusion restriction, and give an alternative estimator for when the large support test fails. Depending on the defendant's race and gender, the probability of wrongful conviction is estimated to range from 63% to 76%, and wrongful acquittal from 4% to 6.5%. These surprising results are due to our sample only including defendants who have reached a final trial, implying substantial evidence against them.

JEL Classification Codes: K14; K41; C14; C25; J15.

1 Introduction

Criminal courts have existed for over a thousand years, yet no reliable method exists to measure their performance. This is not due to apathy: miscarriages of justice have been a source of debate going back to antiquity (Zalman, 2017) and have seen a surge of interest in recent years. Indeed, the popularity of television shows and podcasts

*Department of Economics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. Email: shin.kanaya@essex.ac.uk.

†Department of Economics, Aarhus University, Fuglesangs Alle 4, Aarhus V 8210, Denmark. Email: lntaylor@econ.au.dk. Financial support from the Aarhus University Research Fund (AUFF-26852) is gratefully acknowledged.

chronicling the (in)famous trials of O.J. Simpson, Steven Avery, and the 'Central Park Five', among others, have further thrust the fallibility of courts into the spotlight.

Furthermore, judicial errors are not a matter of mere curiosity. Since its inception in 1992, the Innocence Project has aided in the exoneration of 367 innocent people who had collectively served more than 5000 years in prison.[1] The financial cost of such type I errors is also not small: on average, each prisoner costs the state more than \$33 000 per year to keep behind bars (Mai and Subramanian, 2017). For type II errors, i.e. failing to punish a guilty defendant, the costs fall predominantly on the victims who do not find the justice they deserve, as well as on society more generally through releasing criminals back into the population and eroding belief in the justice system.

In this paper, we estimate these error rates for judges using data on more than five million court cases from Virginia. Our approach involves reframing the problem in the context of misclassified binary choice models, where the misclassification rates are interpreted as type I and type II errors. Specifically, we consider the judge's decision to convict or acquit as a noisy measure of the defendant's true guilt. We provide novel nonparametric identification results for the misclassified binary choice model that admit simple and intuitive estimators. These estimators are of independent interest and have potential applications in other contexts, as discussed later in this section.

The key identification condition requires a continuous 'special regressor' (Lewbel, 1998) that affects the true outcome but is mean independent of the observed (noisy) outcome conditional on the true outcome and a set of regressors. In our context, this conditional mean independence assumption requires a regressor which affects the probability that a defendant is guilty but does not affect the judge's decision. This regressor must also satisfy a large support condition similar to Lewbel (2000b). We provide a novel method to test this large support assumption and develop an alternative estimation scheme for cases that fail this test.

As our special regressor, we use a measure of future criminality constructed from future arrests, convictions, fines, probation, and prison time using machine learning (ML) methods. Applying our test of the large support condition reveals that this condition is satisfied for a wide range of regressor values. We also examine the conditional

[1] The Innocence Project is a 501 nonprofit legal organisation that uses DNA evidence to overturn wrongful convictions.

mean independence assumption by investigating whether the judge's decision affects future criminality. To this end, using a subset of our data where cases are quasi-randomly assigned to judges, we use the conviction tendency of these judges as an instrumental variable (IV) to show that the effect of conviction on future criminality is small and statistically insignificant, adding weight to the validity of the conditional mean independence assumption.

It is important to note that only cases that reach the final trial are included in our analysis; cases settled by a guilty plea or dismissed earlier are removed. Consequently, our results are not generalisable to the entire population of defendants. We answer the more practical question: What is the probability that a judge makes an error in the final trial? Thus, we explicitly condition on reaching the final trial, sidestep the complexities of the earlier stages of the judicial process, and focus on judges' decision-making at the point at which they actually make decisions. While including earlier stages of the judicial process in this analysis would be of great interest, the number of decisions - including those by the defendant during the plea bargaining process - renders the problem almost intractable. Indeed, it is standard practice for justice research to focus on a single stage of the judicial process to constrain the complexity of the problem (see, e.g., Arnold, Dobbie and Hull, 2022; Hoekstra and Sloan, 2022; Park, 2017).

Our baseline results show that the probability of convicting an innocent defendant is much higher than would initially be thought, ranging from 63% to 76%, depending on the defendant's race and gender. However, again we note that these probabilities are conditional on having reached the final trial, i.e. the individual having been arrested and charged, which dramatically increases the probability of conviction relative to the general population. Indeed, the overall conviction rate in the final sample is 82%. This also explains why the probability of acquitting a guilty defendant, ranging from 4% to 6.5%, is lower than might be first expected.

These results may indicate a racial bias against black defendants in terms of the probability of wrongful conviction. Specifically, black defendants face a 10 percentage point higher chance of being convicted when innocent compared to white defendants. Yet, interestingly, black defendants also face a one percentage point higher chance of being acquitted when guilty. This seemingly contradictory result could be explained by bias against black suspects earlier in the judicial process. For example, if police officers require less evidence to arrest black suspects, the average 'strength of evidence' for black defendants at the final trial may be lower than for white defendants. Unfortunately, to explore this hypothesis further, we would need to model the arrest decision using data from the entire population (see Knox, Lowe, and Mummolo, 2020).

Finally, we note that our estimation strategy applies to other decision-making contexts where choices are made with incomplete information but mistakes are unobservable. For example, interviews for prospective employees are designed to gather information about an applicant to avoid hiring an inadequate worker or passing on a suitable one. In some circumstances, it is possible to discover if the wrong person has been hired; however, it is almost always impossible to know if the right person was not hired. This is also true of university admissions, promotions, and lending applications, among many other examples. At the same time, it is important to recognise that our method cannot determine the probability of error for an individual observation, it only provides an average probability conditional on observable characteristics.

1.1 Previous and Recent Literature

This paper contributes to three main literatures. First, we add to work on estimating the prevalence of miscarriages of justice. Almost without exception, the existing research in this area has been restricted to estimating the probability of wrongful conviction - as opposed to wrongful acquittal. Furthermore, this work has, almost exclusively, used data on exonerations (see, e.g., Risinger, 2006; Gross and O'Brien, 2008; Gross, O'Brien, Hu and Kennedy, 2014). However, an exoneration is not equivalent to innocence.

First, the number of exonerations is likely to represent only a small fraction of the total number of false convictions. In many cases, the effort to uncover these miscarriages of justice is not made; the severity of the crime, and hence the punishment, is too low to warrant the use of limited resources. Moreover, even if an investigation is conducted, the evidence required to overturn a previous conviction may not exist. Second, in the majority of cases, exonerations only occur due to misconduct during the arrest or trial.[2]

These limitations of exoneration data are well documented (see, e.g., Acker, 2017), resulting in several attempts to mitigate these shortcomings. A notable example is

[2] National Registry of Exonerations.

Gross *et al.* (2014) who restrict their analysis of exonerations to death row inmates. They reason that by considering a subset of convictions for which the majority of mistakes are identified, their estimate is more accurate. They find the probability of convicting an innocent defendant to be 4.1% but acknowledge this is likely to be a lower bound for the true probability. They also have reservations about extrapolating this to other cases: the judge may spend more time deliberating the evidence or be more likely to err on the side of caution when a person's life is at stake.

To the best of our knowledge, Spencer (2007) is the only work (excluding the present paper) that does not rely on exoneration data and estimates the probability of acquitting a guilty defendant. His method is similar in spirit to ours. By analysing the rate of agreement between a judge and jury, he shows that the probability of a wrongful conviction and of wrongful acquittal can be identified. However, he is transparent regarding the strong assumptions imposed. In particular, judges and juries are assumed to make mistakes at the same rate. Furthermore, the probability of a correct decision from the judge is independent of the probability of a correct decision from the jury in a given trial. This seems unrealistic; for example, in a complex case, the probability of a correct decision is likely to be lower for both judge and jury. In contrast, we only require data on the decision of either a judge or jury - not both.

Finally, Bjerk and Helland (2020) take a different approach and concede that while the exact probability of a false conviction may be beyond reach, differences in the exoneration rate (from DNA evidence) across races can shed light on racial discrepancies in sentencing. They find that the exoneration rate of white defendants for rape cases was less than two-thirds of that for corresponding black defendants. However, their analysis was limited by the small sample size of DNA exonerations.

The second literature to which we contribute concerns racial bias in decision-making. Recent studies (e.g., Canay, Mogstad and Mountjoy, 2022; Hull, 2021; Arnold, Dobbie and Hull, 2022, 2021) have provided insightful discussions on the definition of racial bias, its identification, and its sources (i.e., statistical discrimination or taste-based discrimination). While we remain silent regarding the sources of such bias in our context, as our framework avoids any direct assumption on the judge's decision rule, it is worth noting that the definition of racial bias employed by Arnold, Dobbie and Hull (2022; ADH hereafter) has a direct link to our estimand (see, Section 2.1).

ADH measure racial discrimination in bail decisions using information on judge error rates. These rates are estimated via extrapolation and an identification-at-infinity assumption, which involves considering a hypothetical "supremely lenient" bail judge. Our methodology also employs a form of identification-at-infinity, with an analogous “supremely guilty/innocent defendant”, although perhaps it is more appropriately termed identification-at-the-boundary. A key difference between ADH and the present paper is that ADH consider an observed true outcome (pre-trial misconduct) while our true outcome (factual guilt) is unobservable. Consequently, the methods used to identify the model parameters are quite different: ADH use information on quasi-randomly assigned judges, while we use information on defendants that is unobserved by judges. It is unclear whether each source of identification could be used in the framework of the other. It would also be interesting to investigate whether our method to empirically validate the identification-at-infinity assumption (the large support assumption in our terminology) also applies to the ADH setup.

Finally, this paper also adds to the literature on misclassified binary choice models. The first identification results for this model were given by Hausman, Abrevaya and Scott-Morton (1998); however, their approach was restricted to a parametric model. Lewbel (2000b) extended this to a semiparametric model and used a special regressor for identification. Although we focus on identification of the misclassification rates - unlike Lewbel (2000b) who considers the regression parameters - having obtained the misclassification rates, the regression parameters can then also be identified. Similarly to Lewbel (2000b), we also use a special regressor; however, we provide weaker conditions under which the misclassification rates can be identified and propose a simpler estimator. Indeed, in that paper, he explains that “the estimators provided here are not likely to be very practical, because they involve up to third-order derivatives and repeated applications of nonparametric regression” (pp. 607-608). In contrast, our estimator uses a single nonparametric regression and does not require the estimation of derivatives. Furthermore, we also propose a method to test the crucial large support condition that has long been a thorn in the side of special regressor methods. Examples of other papers which apply special regressor methods include Heckman and Navarro (2007) in a dynamic choice model, Berry and Haile (2014) to estimate demand functions, and Lewbel and Tang (2015) and Khan and Nekipelov (2018) in game-theoretic models.

2 Identification

2.1 Baseline Identification

This section provides details of the general model, the identification strategies, and the required assumptions. Our objects of interest are the type I and type II error probabilities defined as

$$\alpha_1(x) := P\left[Y=1|Y^*=0, X=x\right] \text{ and } \alpha_2(x) := P\left[Y=0|Y^*=1, X=x\right],$$

respectively, where Y and Y^* are binary-valued variables. Y^* denotes the true unobservable outcome, Y is an observed but misclassified version of Y^*, and X represents a vector of observable covariates. In our setting, Y^* denotes whether the defendant is factually guilty ($=1$) or innocent ($=0$), Y indicates whether the defendant was convicted ($=1$) or acquitted ($=0$), and X includes information on both the case and the defendant. Hence, $\alpha_1(x)$ gives the probability of convicting an innocent defendant with characteristics x, and $\alpha_2(x)$ is the corresponding probability of acquitting a guilty defendant.

In connection to ADH, there is a direct link between their measures of discrimination defined in Equations 1 and 2 (pp. 3000) and our $(\alpha_1(x), \alpha_2(x))$ (see also, Arnold, Dobbie and Hull, 2021). In particular, they define discrimination as a weighted sum of $\Delta_0 = E\left[D|D^*=0, R=w\right] - E\left[D|D^*=0, R=b\right]$ and $\Delta_1 = E\left[D|D^*=1, R=w\right] - E\left[D|D^*=1, R=b\right]$, where $R \in (b, w)$ denotes the race (black or white) of the defendant, D is the decision of the judge to release the defendant on bail, and D^* indicates whether there was pre-trial misconduct on the part of the defendant.[3] The goal of the judge is to choose D to match $(1-D^*)$; in our case, the judge attempts to match Y with Y^*. The key difference is that their D^* is observed by the researcher, while our Y^* is not. Furthermore, we condition on additional regressors X (in line with, e.g., Canay, Mogstad and Mountjoy, 2022), while ADH do not. This conditioning has important implications for the interpretability of the discrimination measure; however, we do not wish to focus on issues of discrimination, nor claim that our results imply discrimination of any particular form (see Canay, Mogstad and Mountjoy, 2022, and ADH for a thorough discussion of this complex

[3] Note that in comparison to the definition in ADH, we drop heterogeneity related to the judge for ease of comparison.

issue).

To achieve identification, we assume the availability of an additional (scalar) variable, V, which we term a special regressor. This regressor is assumed to satisfy some conditions distinct from the set of other covariates X. In particular, we assume:

Assumption 1 [Exclusion Restriction] There exists a scalar-valued, continuously distributed variable V which satisfies

$$E\left[Y|Y^{*},X,V\right]=E\left[Y|Y^{*},X\right] \text{ almost surely.}$$

We will use future criminality of the defendant (defined in detail in Section 3.2) as the special regressor, V. In this setup, an error made by the court is given by $(Y-Y^{*})$ and, under Assumption 1, its conditional expectation can be written as

$$E\left[Y-Y^{*}|Y^{*},X,V\right]=E\left[Y-Y^{*}|Y^{*},X\right].$$

This says that, on average, the error depends on the defendant's factual guilt and the characteristics of the case and defendant but not on future criminality. In Section 5, we provide an investigation into the validity of Assumption 1 using the stringency of quasi-randomly assigned judges to cases as an IV for conviction status.

Under this assumption, we can write

$$\begin{aligned}
P\left[Y=1|\left(X,V\right)=x,v\right] &= P\left[Y=1|Y^{*}=0,\left(X,V\right)=x,v\right]P\left[Y^{*}=0|\left(X,V\right)=x,v\right] \\
&\quad +P\left[Y=1|Y^{*}=1,\left(X,V\right)=x,v\right]P\left[Y^{*}=1|\left(X,V\right)=x,v\right] \\
&= \alpha_{1}(x)\left[1-P\left[Y^{*}=1|\left(X,V\right)=x,v\right]\right] \\
&\quad +\left[1-\alpha_{2}(x)\right]P\left[Y^{*}=1|\left(X,V\right)=x,v\right] \\
&= \alpha_{1}(x)+\left[1-\alpha_{1}(x)-\alpha_{2}(x)\right]P\left[Y^{*}=1|\left(X,V\right)=x,v\right]. \qquad (1)
\end{aligned}$$

Note that the objects of interest, $\alpha_1(x)$ and $\alpha_2(x)$, are independent of v. This expression forms the basis of our identification argument; however, one further assumption is required. For clarity of exposition, we assume that the support of $V|X=x$ is a bounded and closed interval $[l_V^x, r_V^x]$ for each $x \in supp\,(X)$, where $supp\,(X)$ denotes the support of the random vector X.[4]

[4] Note that, with only slight modifications, all subsequent results carry over to cases that allow for an unbounded or (semi-) open interval, including $(-\infty,\infty)$.

Assumption 2 [Large Support Condition] For each $x \in supp\,(X)$,

$$\lim_{v \to l_V^x} P\left[Y^* = 1 | \left(X, V\right) = (x, v)\right] = 0, \tag{2}$$

$$\lim_{v \to r_V^x} P\left[Y^* = 1 | \left(X, V\right) = (x, v)\right] = 1. \tag{3}$$

Assumption 2 states that being factually guilty or not can be perfectly predicted by future criminality in its tail region. In other words, the support of V is sufficiently large: $[l_U^x, r_U^x] \subseteq [l_V^x, r_V^x]$, where $[l_U^x, r_U^x]$ is the support of $U|X = x$.[5] This assumption can be viewed similarly to the 'supremely lenient judge' assumption of ADH.

From equation (1) and Assumption 2, we obtain

$$\lim_{v \to l_V^x} P\left[Y = 1 | \left(X, V\right) = (x, v)\right] = \alpha_1(x), \tag{4}$$

$$\lim_{v \to r_V^x} P\left[Y = 1 | \left(X, V\right) = (x, v)\right] = 1 - \alpha_2(x), \tag{5}$$

which establish the identification of $\alpha_1(x)$ and $\alpha_2(x)$, respectively. When the support of $V|X = x$ is $(-\infty, \infty)$, this type of identification is typically referred to as 'identification-at-infinity'. However, in our setting, V is calculated as a probability and therefore has known support of $[0, 1]$; thus, this problem concerns boundary estimation, so is more akin to regression discontinuity design, and may be more appropriately called 'identification-at-the-boundary'.

2.2 Testing the Large Support Condition

It is clear that the large support condition is critical to achieving identification. Thus, it is important to provide a method to empirically check its validity. To this end, we introduce an additional assumption:

Assumption 3 [Single-Index Structure] The true outcome Y^* and regressors (X, V) are related through

$$Y^* = \mathcal{I}\left(V + h(X) - U \geq 0\right), \tag{6}$$

[5] Assumption 2 is written using limit notation such that the conditions hold for unbounded or open support settings. This reduces to $P\left[Y^* = 1 | \left(X, V\right) = (x, l_V^x)\right] = 0$ and $P\left[Y^* = 1 | \left(X, V\right) = (x, r_V^x)\right] = 1$ when the support of $V|X = x$ is $[l_V^x, r_V^x]$.

where $\mathcal{I}(\cdot)$ is the indicator function, $h(\cdot)$ is an unknown scalar-valued function on the support of X, U is an unobservable random variable with $U \perp V|X$, and $U|X = x$ is continuously distributed for each x, i.e. the conditional cumulative distribution function (CDF) of U, $F_{U|X}(u|x)$, has a corresponding density $f_{U|X}(u|x)$.

The conditional independence condition $U \perp V|X$ in Assumption 3 resembles that of Lewbel (2000a, Assumption A2). This, together with the single index structure of Y^* in equation (6), leads to the following expression for the 'conditional predictive probability' (CPP):

$$P[Y^* = 1|(X, V) = (x, v)] = F_{U|X}(v + h(x)|x). \tag{7}$$

While Assumption 3 may look restrictive, it does not impose any significant restriction on the functional form of the CPP except for monotonicity in v. That is, any CPP that is monotone in v can be represented by the model in Assumption 3 under mild regularity conditions (cf. Theorem 3 of Magnac and Maurin, 2007, who give a representation result for monotone binary choice models). In Appendix A, we provide a representation theorem for the CPP with some further discussion. Note that this single-index model is not a structural model, i.e. it does not attempt to explain a defendant's criminal behaviour. It is merely a tool for the researcher to predict such behaviour retrospectively. This stands in contrast to previous work that uses similar single-index specifications and conditional independence assumptions to create structural models (see, e.g., Berry and Haile, 2014), where careful consideration must be given to the underlying behavioural mechanisms that could result in such a model.

To construct our test, first note that $F_{U|X}(r_V^x + h(x)|x) = 1$ is a necessary and sufficient condition for Equation 2 to hold; equally $F_{U|X}(l_V^x + h(x)|x) = 0$ is a necessary and sufficient condition for Equation 5 to hold. Under Assumption 1-3, the partial derivative of equation (1) with respect to v is

$$\frac{\partial}{\partial v} E[Y|V = v, X = x] = [1 - \alpha_1(x) - \alpha_2(x)] f_{U|X}(v + h(x)|x). \tag{8}$$

Note that the left-hand side of equation (8) can be directly estimated from the data. Furthermore, for a given x, the right-hand side is a constant multiple of $f_{U|X}(v + h(x)|x)$. Thus, fixing x, it is possible to evaluate this derivative in the interval $[l_V^x, r_V^x]$ to determine whether the tail condition is satisfied for a given x. If the derivative is zero in the upper limit of V, this suggests $F_{U|X}(r_V^x + h(x)|x) = 1$, providing that $f_{U|X}(\cdot|x)$ has no zero-probability intervals in the interior of its support that are outside the support of V. Equally, a zero derivative at the lower limit of V indicates $F_{U|X}(l_V^x + h(x)|x) = 0$. Thus, the validity of the large support assumption can be checked for each tail condition and for each point of interest x. We implement this test in our empirical application in Section 6.

2.3 Relaxation of the Large Support Condition

While Assumption 2 appears to be satisfied in our empirical application (see Section 6), in many empirical settings, this may not be the case. As such, it is worthwhile to pursue alternative identification mechanisms which do not require equations (2) and (3) to hold simultaneously. Without loss of generality, we proceed without the lower-bound condition (2) and impose the following assumptions:

Assumption 2′ [Alternative Large Support Condition] For each $x \in supp\,(X)$,

$$\lim_{v \to r_V^x} P\left[Y^* = 1|\,(X,V) = (x,v)\right] \;=\; 1. \tag{9}$$

Assumption 4 [Mode-Median Coincidence/Limited Predictability] The conditional CDF $F_{U|X}(\cdot|x)$ is differentiable on the entire support of $U|X = x$ and has derivative $f_{U|X}(\cdot|x)$. There exists a unique global maximum point (conditional mode) m_U^x of $f_{U|X}(\cdot|x)$ on $[l_V^x + h(x), r_U^x]$ which coincides with the conditional median of $U|X = x$, where r_U^x is the upper limit of the support of $U|X = x$.

Assumption 2′ is a weakening of Assumption 2 in that it removes the lower tail condition but maintains the upper (i.e. it is only supposed that $r_U^x \leq r_V^x$).

Under Assumptions 1 and 2′ it is possible to identify $\alpha_2(x)$ but not $\alpha_1(x)$. However, Assumptions 3 and 4 can be used to recover $\alpha_1(x)$. The restriction involving the conditional mode may look unusual for latent-variable discrete choice models; nonetheless, this assumption is satisfied by commonly-used parametric distributions. Its simplest sufficient condition is that $U|X = x$ is symmetric and unimodal, as in the case of the Gaussian or logistic distributions; however, it does not exclude non-symmetric distributions.[6]

[6] Note that the maximum point m_U^x need not necessarily be the mode of $U|X = x$. That is, the

We interpret Assumption 4 as a limited predictability condition in the following way. From the form of the CPP in equation (7), the median value of $U|X = x$ occurs where the probability of the defendant being guilty is 0.5. Since we require $\text{Mode}\,[U|X = x] = \text{Median}\,[U|X = x]$, it must be that there is a significant proportion of defendants who are as likely to be guilty as they are to be innocent and, consequently, whose guilt is difficult for the researcher to predict. Importantly, note that this is not required to hold in the population as a whole. We only require, for a given point of interest x, some value of V for which this holds.

It is also possible to interpret Assumption 4 as a type of location normalisation: Manski (1988) imposes a location normalisation through a conditional median restriction to identify $h(x) = x'\beta$. In the present context, his assumption corresponds to $\text{Median}\,[U|X] = 0$. While $\text{Mode}\,[U|X] = 0$ can play the same role, it is important to note that Manski (1988) considers an observable binary outcome. If Y^* were observable in our setting, either $\text{Mode}\,[U|X] = 0$ or $\text{Median}\,[U|X] = 0$ could be used to identify $h(x)$.[7,8] In this respect, Assumption 4 is stronger than necessary when Y^* is observable. Theorem 2 in Appendix A gives a representation result for the CPP and clarifies that indeed $\text{Mode}\,[U|X] = \text{Median}\,[U|X]$ imposes more structure on the CPP in equation (7) than either $\text{Mode}\,[U|X] = 0$ or $\text{Median}\,[U|X] = 0$. However, it appears that when Y^* is unobservable, some additional restriction, such as Assumption 4, must be imposed for identification.

We now illustrate how Assumption 4 can be used to restore the identification of $\alpha_1(x)$ when Assumption $2'$ holds but Assumption 2 does not, i.e. when only the upper tail condition is satisfied. Recall that

$$P\left[Y = 1 | \left(X, V\right) = (x, v)\right] = \alpha_1(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right] F_{U|X}\left(v + h(x)|x\right).$$

Taking the partial derivative with respect to v gives

true mode may exist outside of $[l_V^x + h(x), r_U^x]$. We simply require that the unique maximum point inside $[l_V^x + h(x), r_U^x]$ is equal to the median; nonetheless, we maintain the mode interpretation for ease of understanding.

[7] It is not necessary to assume that these conditional measures are equal to 0. It is possible to use any known number $c_x \in \mathbb{R}$ for each x normalisation instead (see Theorem 2 in Appendix A).

[8] In this case, an additional condition would be required for identification: for the former mode condition, there must exist some v such that $P\left[Y^* = 1|(X, V) = (x, v)\right] = 1/2$ for each x; and for the latter median condition, $(\partial/\partial v)P\left[Y^* = 1|(X, V) = (x, v)\right]$ must have a unique maximiser v in the support of V for each x (see Theorem 2 in Appendix A).

$$\frac{\partial}{\partial v}P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]=\left[1-\alpha_1(x)-\alpha_2(x)\right]f_{U|X}\left(v+h(x)|x\right).$$

Since the right-hand side is a constant multiple of $f_{U|X}\left(v+h(x)|x\right)$ for a given x, if $\alpha_1(x)+\alpha_2(x)<1$, we can define

$$\begin{aligned}\bar{v}(x) &:= \mathrm{argmax}_{v\in\left[l_V^x,r_V^x\right]}\frac{\partial}{\partial v}P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]\\ &= \mathrm{argmax}_{v\in\left[l_V^x,r_V^x\right]}f_{U|X}\left(v+h(x)|x\right).\end{aligned}\tag{10}$$

Note that $\frac{\partial}{\partial v}P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]$ is identified directly from the data; thus, it is straightforward to estimate $\bar{v}(x)$. The restriction $\alpha_1(x)+\alpha_2(x)<1$ is what Hausman *et al.* (1998) call the monotonicity condition and is standard in the literature on misclassified binary variables. In our empirical setting, this states that the court's ruling is informative of the guilt of the defendant. If this did not hold, the court would make fewer mistakes if all those who were convicted were acquitted instead, and all those originally acquitted were now convicted.

Here $\left[\bar{v}(x)+h(x)\right]$ is the unique maximiser of $f_{U|X}\left(v+h(x)|x\right)$, hence, it constitutes the modal point of $U|X$. Moreover, under Assumption 4, $\left[\bar{v}(x)+h(x)\right]$ is the median of $U|X=x$, so $F_{U|X}\left(\bar{v}(x)+h(x)|x\right)=1/2$. Armed with this, we can write

$$\begin{aligned}P\left[Y=1|\left(X,V\right)=\left(x,\bar{v}(x)\right)\right] &= \alpha_1(x)+\left[1-\alpha_1(x)-\alpha_2(x)\right]F_{U|X}\left(\bar{v}+h(x)|x\right)\\ & \quad\ \alpha_1(x)+\left[1-\alpha_1(x)-\alpha_2(x)\right]/2\\ &= \left[1+\alpha_1(x)-\alpha_2(x)\right]/2.\end{aligned}$$

After obtaining $\alpha_2(x)$ using equation (9), which holds under Assumption $2'$, we can then obtain $\alpha_1(x)$ since the left-hand-side is directly identified from the data. We summarise this result in the following theorem.

Theorem 1 Suppose that Assumptions 1, $2'$, 3, and 4 hold. If $\alpha_1(x)+\alpha_2(x)<1$ for each $x\in supp\left(X\right)$, then $\alpha_1(x)$ and $\alpha_2(x)$ are identified as

$$\begin{aligned}\alpha_1(x) &= 2P\left[Y=1|\left(X,V\right)=\left(x,\bar{v}(x)\right)\right]+\alpha_2(x)-1,\\ \alpha_2(x) &= 1-\lim_{v\to r_V^x}P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right],\end{aligned}$$

where $\bar{v}(x)$ is defined in equation (10).

In principle, estimators for $\alpha_1(x)$ and $\alpha_2(x)$ can be constructed using empirical analogues of the expressions in Theorem 1. However, the following integral-based formula (derived in Appendix A) may be more practical:

$$\begin{aligned}\alpha_2(x) &= 1 - P\left[Y = 1 | X = x\right] \\ &\quad - \int_{l_V^x}^{r_V^x} \left(\partial / \partial v\right) P\left[Y = 1 | \left(X, V\right) = (x, v)\right] F_{V|X}(v|x) dv. \qquad (11)\end{aligned}$$

An estimator based on this object is likely to be more robust and allow an easier analysis of its asymptotic properties (cf. Goh, 2018).

3 Institutional Setting and Data

3.1 Institutional Setting

The Virginia trial court system includes four levels of court: the Supreme Court, the Court of Appeals, the Circuit Courts, and the General District Courts. As indicated in Appendix B, Figure B.1, following arrest, cases enter Virginia's judicial system through magistrates who determine if there is sufficient evidence to warrant a charge and, if so, at which level of court the case should be heard. Juvenile and Domestic Relations District Courts preside over cases involving juveniles and family relationships. General District Courts make rulings on civil cases, traffic offences, and other minor crimes; they only hold preliminary hearings for felony cases before transferring them to a Circuit Court. Circuit Courts are the highest trial courts with general jurisdiction in Virginia; they hear more serious crimes and appeals from both the General and Juvenile District Courts. The Court of Appeals handles appeals from the Circuit Courts, while the Supreme Court reviews decisions by the Court of Appeals and lower courts.

In this empirical study, we use data from Virginia's 32 General District Courts. There is a General District Court in every city and county in Virginia and these courts decide all infractions and misdemeanours. Infractions carry no prison sentence and a maximum fine of \$250, while misdemeanours can be punished by a maximum prison sentence of one year, a fine of up to \$2500, or both. A judge also has the option to impose probation or a suspended sentence.

General District Courts do not conduct jury trials - all cases are heard by a judge. The judges of Virginia's district courts are elected by a majority vote of each house of the General Assembly for terms of six years. In the data, we only observe individuals who have already been arrested and charged with a crime. After being charged, a judge holds a preliminary hearing to decide whether to dismiss the case before going to a full trial (this judge is not necessarily the same judge who hears the final trial). If the case is not dismissed before the trial, the defendant has the option to plead guilty and receive a lesser sentence. Failing this, the trial ensues and if the defendant is found guilty, the judge determines the penalty. The Code of Virginia defines criminal offences and the range of penalties to be imposed. Throughout this paper, we are only interested in the conviction decision, not the sentencing.

General District Courts also decide civil cases. In a civil case, the court is asked whether a defendant is guilty *on the balance of probability*; thus, the judge need only be 50% sure of guilt in order to convict. This is in contrast to criminal cases where the court is asked whether a defendant is guilty *beyond reasonable doubt.* This 'doubt' is the type I error we are interested in, i.e. the probability of convicting an innocent defendant. As such, our analysis considers only criminal cases. As stated previously, General District Courts also hold preliminary hearings in felony cases (defined as any charge punishable by more than one year in prison). Since these are only preliminary hearings, we do not consider them in our analysis. The choice to limit our investigation to General District Courts and hence only to lower level crimes is to help ensure the validity of the zero conditional mean assumption; this restriction is discussed in more detail in Section 3.2.

3.2 The Special Regressor

Due to its critical role in our analysis, before outlining the data, we discuss the choice of future criminal behaviour as the special regressor; precise details of its construction are deferred to Section 3.3. Section 6 provides evidence that this variable is highly correlated with the true guilt of the defendant and satisfies the large support condition for a range of control variable values. Intuitively, if a judge were informed, for example, that a defendant will be convicted for the same type of crime many times in the future, this would provide substantial information regarding the defendant's likelihood of guilt for the current crime. However, since future criminality does not materialise until after the trial, there is no means for it to affect the court proceedings. Nonetheless, there are still two concerns.

First, conditional on true guilt and a set of controls, future criminality must be unrelated to unobservables which affect the probability of misclassification. Focussing on false acquittal, essentially, our estimation strategy uses the release rate of individuals with very high future criminality as the false acquittal rate. To extrapolate this rate to individuals with lower levels of future criminality requires that all individuals share the same misclassification rate conditional on observable characteristics and true guilt. We believe that by nonparametrically controlling for key variables such as race, gender, and previous criminality, we close off many potential channels through which unobserved characteristics could lead to different misclassification rates. Moreover, if an unobservable variable affects the probability of a miscarriage of justice through a channel other than true guilt, this variable must represent a bias in the judge's decision; consequently, it is unlikely to be related to the future criminal behaviour of the defendant.

If there is still concern regarding a particular unobserved variable, the interpretation of the misclassification rate must be altered; the misclassification rate becomes specific to the group of individuals who have high future criminality. For example, if those with visible gang tattoos (which are not observed by the researcher) are more likely to have high future criminality and also have a different wrongful acquittal rate (conditional on control variables and true guilt), then the misclassification rate becomes specific to those with visible gang tattoos.

A second issue is whether the judge's decision can affect future criminality. If this is the case, future criminal behaviour and the court ruling will not be mean independent. To mitigate concerns of this nature, the sample is first restricted to infractions and misdemeanours to reduce the potential effect of a conviction. The mean prison sentence for those convicted in our sample is only 11 days, and the median is no prison time at all. Although this removes some of the more interesting offences, it is possible that the less serious cases are likely to be subject to more frequent miscarriages of justice: the conveyor belt of defendants passing through the justice system results in judges and lawyers giving less time to each case. Notwithstanding this data restriction, in Section 5, we go on to formally test whether the judge's decision affects future criminality.

3.3 Sample and Variable Construction

The data cover all arrests for which charges were filed for the years 2009-2020, where the unit of observation is a single charge. Each observation provides information on the defendant's gender, race, and address, as well as details of the charge and the outcome of the criminal proceedings. The initial dataset contains more than 20 million observations.

The special regressor, future criminality, is constructed from several measures of future criminal behaviour which must be individually calculated from the sample. To this end, a unique identifier for each individual based on their full name, gender, race, and day and month of birth is used to create the following variables: the number of arrests (including parole violations), the number of arrests for the same type of crime as the defendant is currently on trial,[9] the number of convictions, the number of convictions for the same type of crime, the dollar amount of fines charged, the number of days sentenced to prison, the number of days for suspended sentences, and the number of days sentenced to probation. Since data are only observed until the end of 2020, these measures of future criminal behaviour are averaged over the number of years between the date of the current trial (or the date of release if the defendant was convicted) and the end of the sample period.[10] We also construct these same eight measures for previous criminality using an analogous procedure. Having calculated these past and future variables, the first and last year of the data are dropped as 'burn-in periods' (losing 13% of observations).

At this stage, the sample is further restricted in several ways. We remove all individuals who do not reside in Virginia (22% loss of observations). Only infractions and misdemeanours are considered (8% loss of observations). This limits the severity of the potential punishment, mitigating the impact of the court decision on future criminal behaviour. Recall that each observation represents a single charge; thus, for trials that have multiple charges, one trial will produce several observations in our data. To avoid issues of dependence, all observations for which the individual faces multiple charges in a single trial are removed (28% loss of observations). We also restrict our attention to the five most common categories of crime in the dataset:

[9] The crime type is defined by the Virginia Crime Codes Statute Order. There are 751 unique crime types in the sample.

[10] Unfortunately, we do not have information regarding parole, and thus cannot adjust our measures to account for this.

traffic violations, crimes related to drugs, violent crimes, property-related crimes, and court violations to ensure a degree of homogeneity in our sample (14% loss of observations).

As discussed in the Introduction, we are only interested in cases that reach the final trial. Therefore, we remove cases that are dismissed prior to trial or are settled by a guilty plea (31% loss of observations). It is important to note that in the case of plea bargaining, the decision to plead guilty is made by the defendant, not the judge. If a defendant pleads guilty, they are always convicted. This means that guilty plea cases are not relevant for our question, since a mistake would never be made if the defendant was actually guilty and would always be made if they were innocent. After applying these restrictions, our sample contains 5.6 million observations.

Each of the eight aforementioned measures of future criminality is a viable choice for the special regressor. However, our method requires only a single variable. To this end, we use ML techniques to combine the measures of future criminality to create a single IV which captures as much information as possible (see, e.g., Belloni, Chen, Chernozhukov and Hansen, 2012; Hartford, Lewis, Leyton-Brown and Taddy, 2017). In particular, the gradient boosted regression tree (GBRT) method of Friedman (2001) is used.[11] Ideally, the outcome of interest in this first stage would be true guilt, but this is unobservable; instead, we use conviction. Under the conditional mean independence assumption, the best linear predictor of guilt is also the best linear predictor of conviction. However, this is no longer true when using nonlinear prediction techniques, as we do. Nonetheless, the approach still yields a strong predictor of guilt, evidenced in Section 6. Recall that the goal is not to determine the true conditional mean of guilt, only to create a measure which is highly predictive of this guilt in the tail regions.

Note that constructing V by examining several methods and selecting one of them should not be misunderstood as a form of data snooping or p-hacking. In a p-hacking-like procedure, a researcher implements several methods and selects the method which leads to some 'favourable final result' (in our context, this 'final result' would be the type I and type II error probability estimates). However, our GBRT

[11] GBRT is an ensemble algorithm that combines many regression trees to approximate a conditional expectation function (see Ch. 29 of Hansen, 2022, for a concise discussion). Other approaches may also be used; however, we found that the special regressor constructed using GBRT gave better results in our empirical checks of the identification assumptions (see Sections 5 and 6) than either random forests or logit.

method is not selected to obtain a particular estimation result but to best satisfy the empirical checks of the identification assumptions. Data snooping is result-driven; in contrast, our procedure is "assumption-driven" and, consequently, does not lead to systematically distorted estimation results.[12]

The data are evenly randomly split into a training set and a hold-out set. With the training set, 5-fold cross-validation and a grid search are used to find optimal choices for tuning parameters. Specifically, we choose: the shrinkage parameter, number of trees, tree depth, minimum number of observations in the terminal nodes, and the fraction of the sample randomly chosen to propose the next tree. The optimal values are then used to estimate the GBRT on the training dataset.[13] The measure of future criminality is given as the predicted conviction probability from this regression tree on the hold-out data. This process is repeated but with the roles of the training and hold-out data reversed. This cross-fitting procedure (Chernozhukov *et al.*, 2018) preserves the full sample size while avoiding the overfitting bias associated with ML techniques. Appendix B, Figure B.2, gives a plot of the distribution of this variable for each race-gender group. In Appendix A, we provide a detailed discussion of the implications of using a special regressor constructed from a regression-type model, and how this impacts the likelihood of satisfying Assumption 3 or $3'$.

One further control variable is constructed from the data. To avoid complicated nonparametric fixed-effects estimators, a ZIP code pseudo-fixed-effect is calculated by taking a leave-one-out average conviction rate for each ZIP code. There are 892 ZIP code areas in Virginia with a mean population of 9,325 and a median of 2,940. This variable aims to control for the unobserved heterogeneity across neighbourhoods.

Indeed, Altonji and Mansfield (2018) give credibility to this idea. Translated into our context, they explain that if the decision to convict a defendant is based on both individual and neighbourhood characteristics and that individuals choose their neighbourhood endogenously, a bias can arise. However, under certain assumptions, controlling for means of observable individual factors at the neighbourhood level can "absorb all of the between-group variation in both observable and unobservable individual inputs" (p. 2903). To achieve this perfect control of unobserved neighbourhood

[12] We also note that our methodology is immune to so-called algorithm bias: ML algorithms may misclassify defendants based on their characteristics (see, e.g., Mehrabi et al., 2021). However, whether V contains such misclassification is irrelevant for our final error probability estimates insofar as V satisfies the identification assumptions.

[13] We use the 'caret' package in R for this cross-validation and estimation.

effects, the utility function of individuals who choose in which neighbourhoods to live must be additively separable in the amenities of the neighbourhood. Furthermore, the number of amenities that have an effect on court proceedings must not be larger than the number of neighbourhood averages included. That is, for full control of unobserved neighbourhood effects, we require judges to use a one-dimensional measure of neighbourhood quality in their decision to convict or acquit. As a robustness check, we repeated our empirical analysis using an additional ZIP code pseudo-fixed-effect based on the per-capita number of arrests and found little difference in results, giving weight to the idea of judges using only a one-dimensional measure of neighbourhood quality.

We close this section with a final remark. Since the sample consists of individuals who have been arrested and subsequently charged, the analysis is conducted conditional on this fact. That is, the following objects are estimated

$$\begin{aligned} \alpha_1(x) &= P\left[Y = 1 \middle| Y^* = 0, X = x, A = 1\right], \\ \alpha_2(x) &= P\left[Y = 0 \middle| Y^* = 1, X = x, A = 1\right], \end{aligned}$$

where A denotes whether the individual has been arrested and charged $(= 1)$ or not $(= 0)$. Throughout, the notational dependence on A is dropped for convenience; however, the distinction should not be forgotten. We estimate the likelihood of a defendant - who has already been arrested and charged - being wrongfully convicted or acquitted. These estimates will be very different for a defendant relative to a member of the general public.

3.4 Descriptive Statistics

Table 1 reports the mean and standard deviation of each variable for defendants based on conviction status. There is very little race or gender difference in those who are acquitted versus convicted. Males make up the majority of the sample, and given that the population of Virginia is approximately 56% white (non-Hispanic), it is unsurprising that the sample is predominantly white. It is interesting to note that 29% of judges in Virginia are black (George and Yoon, 2017), corresponding almost exactly with the proportion of black defendants. Thus, it is equally likely that a black defendant faces a white judge, as it is a white defendant faces a black judge.

Future criminality is only slightly higher for those who are convicted relative to those who are acquitted. In addition, there appears to be little difference in previous criminality or the neighbourhood effect across convicted and acquitted defendants. This is likely a reflection of the types of crimes which lead to a conviction; a higher fraction of crimes resulting in conviction are infractions rather than misdemeanours.

Table 1: Descriptive Statistics

	Convicted		Acquitted	
	Mean	St. Dev.	Mean	St. Dev.
Future Criminality	0.82	0.05	0.81	0.06
Previous Criminality	0.82	0.06	0.80	0.06
Neighborhood Effect	0.81	0.06	0.79	0.06
Male	0.59	0.49	0.58	0.49
Black	0.30	0.46	0.32	0.47
Infraction	0.82	0.38	0.62	0.48
Drug Crime	0.02	0.15	0.06	0.23
Property Crime	0.01	0.09	0.03	0.16
Traffic Crime	0.94	0.24	0.80	0.40
Violent Crime	0.02	0.15	0.09	0.29
Court Crime	0.01	0.08	0.02	0.15
Observations	4,579,733		1,006,959	
Percentage	82%		18%	

Notes: This table displays means and standard deviations for the final sample of defendants (selected according to the criteria given in Section 3.3) used to estimate the misclassification rates presented in Section 4.

4 Results

We use a local linear likelihood estimator with a logistic link function to estimate the nonparametric functions.[14] Frölich (2006) showed in a series of Monte Carlo simulations for a binary choice model that local linear logit is substantially more precise than the Nadaraya-Watson estimator, the local linear kernel estimator, the semiparametric estimator of Klein and Spady (1993), and parametric logit. Furthermore, Fan, Heckman, and Wand (1995) showed that local linear logit has desirable bias proper-

[14] Smoothing uses the tricube kernel and nearest neighbours, where the number of neighbors is 10% of the sample size; results are qualitatively similar for 5% and 20% (see the Online Appendix).

ties at the boundaries, which is important as our approach is based on an evaluation of $E[Y|X, V]$ at the boundaries of V. Since V is a probability, the endpoints are 0 and 1, so we need not estimate them when evaluating $E[Y|X, V]$, we can simply plug in 0 and 1; using estimated endpoints would affect the asymptotic properties of the estimator. It is worth noting that for each race and gender pair, the endpoints of V in the sample are within 0.03 of 0 and 1. Furthermore, Figures 6.1 and 6.2 in Section 6, show $\partial E[Y|X, V = v]/\partial V \approx 0$ for v in a wide interval around 0 and 1; thus, there is little sensitivity of $E[Y|X, V = v]$ to v in these areas.

The control variables include the race, gender, and previous criminality of the defendant, the neighbourhood effect, whether the crime is an infraction, and the type of crime (five categories). When evaluating the misclassification probabilities, we set the neighbourhood effect at its mean, fix the crime to be a traffic infraction (the most common crime), and plot results for each race and gender group over a range of previous criminality values (from the 0.1 to the 99.9 percentile). In Section 6, we show that at these values of the regressors, the large support assumption is satisfied.

Figure 4.1 plots the probability of being incorrectly convicted after reaching the final trial for each race-gender group as a function of previous criminality, together with a 95% pointwise confidence band based on a smoothed bootstrap; Claeskens and Van Keilegom (2003) show the validity of this bootstrap for local likelihood estimators.[15] Note that the bootstrap procedure also takes into account the construction of V.

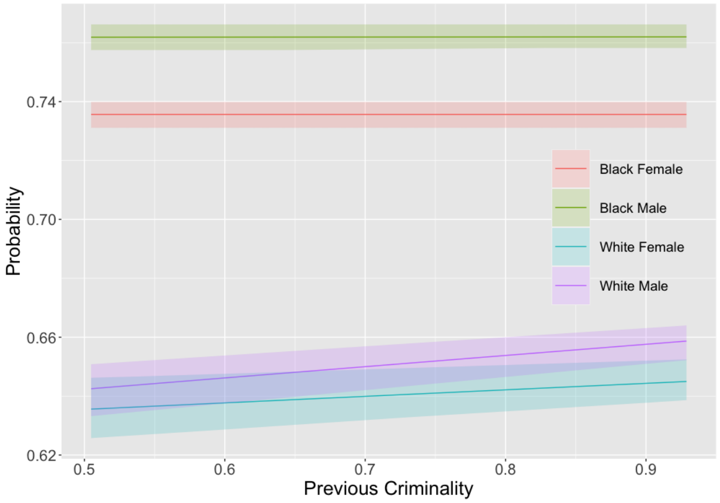
First, the likelihood of convicting an innocent black defendant is higher than that for an innocent white defendant, regardless of gender. In particular, an innocent black male who has reached a final trial faces a worryingly high probability of being convicted. Indeed, each demographic group has a high chance of conviction when innocent. However, recall that this analysis is conducted conditional on being arrested and charged for the crime. In order for an individual to be arrested and charged, there must be compelling evidence against them. Thus, these estimates will be very different from estimates for a member of the general population.

A bias against men relative to women also exists that is consistent across race

[15] As explained in Section 3.3, our procedure selects the best estimator for V based on assumption checks. While the uncertainty associated with this should be accounted for, this is not an easy task and we do not pursue it here. Our estimation process has an issue analogous to the post-model-selection inference problem, and its solution is known only for specific cases; see, e.g., Bachoc et *al.*, 2020, for relatively simple regression cases.

groups. At the mean of previous arrests, innocent white men face a conviction probability of 65.4%, compared to 64.2% for white females. Similarly, innocent black males have a 76.2% probability of conviction, in contrast to 73.6% for black females. We also see that previous criminality has only a slight positive effect on conviction probability; in fact, the 95% confidence band for each group contains a constant effect.

Figure 4.1: Probability of Convicting an Innocent Defendant

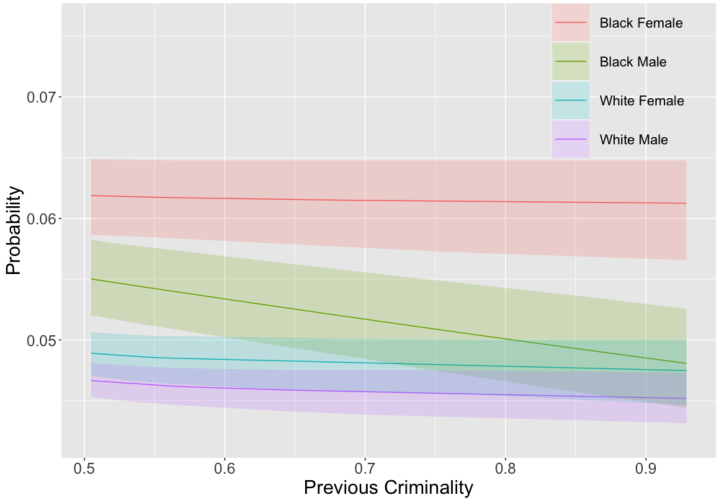


Notes: This figure plots estimates for the probability of convicting an innocent defendant using the identification scheme in Section 2.1. Estimation uses a local logit with 10% nearest neighbours on the sample of defendants (separated by race and gender) as detailed in Section 3.3. Control variables include the neighbourhood effect, the type of crime and whether it is an infraction, and the previous criminality of the defendant. The neighbourhood effect is set at its mean and the crime is set as a traffic infraction. 95% pointwise confidence bands are also given.

In Figure 4.2, we plot estimates for the probability of acquitting a guilty defendant. Again, there is a bias against men in favour of women, although the difference is small. At the average value of previous criminality, the probability of acquittal for a guilty black male is 5%, compared to 6.1% for black females. Likewise, guilty white males face a 4.5% chance of being acquitted, in comparison to a 4.8% likelihood for white females. Taken together with the results of Figure 4.1, it appears the threshold for convicting a woman is lower than that for a man. Finally, it should not be surprising that the probability of wrongful acquittal decreases as the number of previous arrests increases, which reflects the greater likelihood of convicting a defendant with a particularly criminal past.

Interestingly, despite the probability of wrongful conviction being higher for black defendants, the probability of wrongful *acquittal* is generally higher for black defendants as well. Furthermore, this pattern is consistent across genders. One potential explanation for this finding relates to the adequacy of the model. It may be that while the conditional mean independence assumption holds over the whole sample, it may not hold for each race individually. If conviction has a positive effect on future criminality for white defendants, this would result in an underestimation of the probability of wrongful acquittal for this group. However, our robustness checks in the Online Appendix refute this premise. We give results of the IV regression discussed in Section 5 estimated on the subsample of black and white defendants separately. The results are similar in both cases, indicating the conditional mean independence assumption holds for both races.

Figure 4.2: Probability of Acquitting a Guilty Defendant



Notes: This figure plots estimates for the probability of acquitting a guilty defendant using the identification scheme in Section 2.1. Estimation uses a local logit with nearest neighbours on the sample of defendants (separated by race and gender) as detailed in Section 3.3. Control variables include the neighbourhood effect, the type of crime and whether it is an infraction, and the previous criminality of the defendant. The neighbourhood effect is set at its mean and the crime is set as a traffic infraction. 95% pointwise confidence bands are also given.

This result could instead be explained by black defendants having more procedural flaws in their cases than white defendants. For example, if police officers violate the constitutional rights of black suspects more often than white suspects, such as through illegal searches, the judge is obliged to dismiss more cases against black suspects despite perhaps believing the defendant to be guilty. Thus, although our method detects incorrect decisions, it may not necessarily be due to judicial errors, but rather policing errors.

5 Validity of the Conditional Mean Independence Assumption

There is an abundance of previous work on the effects of incarceration on future criminality that is somewhat contradictory. For example, Mueller-Smith (2015) found that incarceration increases recidivism, while Mitchell, Cochran, Mears and Bales (2017) found no effect, and Bhuller, Dahl, Løken and Mogstad (2020) even suggested that prison time reduces future criminality. In contrast, there is relatively little research on how conviction without incarceration affects future criminal behaviour - a more relevant question in our context. And, as with the work on incarceration, this research is contradictory: Ventura and Davis (2005) found that convictions reduce the likelihood of recidivism, while Chiricos, Barrick, Bales and Bontrager (2007) showed the opposite effect.

Due to these contradictory findings, it is difficult to use them to defend the conditional mean independence assumption in our setting. Consequently, to test this assumption, we use the conviction tendency of quasi-randomly assigned judges as an IV to uncover the causal effect of conviction on future criminality.

5.1 IV Institutional Setting and Data

We restrict our analysis to Chesterfield County General District Court, where we have confirmation from a personal correspondence with the court clerk that judges are randomly assigned to cases, with the exception of probation violation cases, which are assigned to the judge who handled the original case. As such, to maintain the random assignment, all probation violation cases are removed from this IV analysis. Additionally, due to the specific days and times that traffic violations are heard at this court, we must account for this when calculating our measure of judge stringency.

Although we were unable to obtain confirmation of random assignment for other courts, the sample size is still large (85,865) and it is not unusual to restrict analyses of this type to a single court to increase the homogeneity of cases (see, e.g., Mueller-Smith, 2015). As with other district courts in Virginia, Chesterfield County judges are elected by the General Assembly and serve six-year terms. At any one time, Chesterfield County General District Court has five judges (including one chief judge). The proceeding analysis follows standard practices in the 'judges-design' IV literature (see, e.g., Bhuller *et al.*, 2020).

We calculate stringency for each judge in each year to allow for changes over time. Specifically, we use the leave-one-out average conviction rate for each judge in each year, then residualise this average for the year the case was filed and for whether the crime is a traffic violation. That is, we estimate the following regression

$$C_i \;=\; \alpha_{t(i)} D_{t(i)} + \beta W_i + \gamma_{t(i)} D_{t(i)} W_i + \epsilon_i, \tag{12}$$

where C_i denotes whether case i resulted in a conviction; $D_{t(i)}$ is a set of dummy variables for the year, t, in which case i was filed (with the corresponding coefficients $\alpha_{t(i)}$), W_i denotes whether case i was a traffic violation, and $\gamma_{t(i)}$ are the coefficients for the interaction between year and traffic violation. The leave-one-out residualized stringency measure is then constructed as

$$Z_i \;=\; \frac{1}{n_j - 1} \sum_{l \neq i} \mathcal{I}\{j(i) = j(l)\} \hat{\epsilon}_l,$$

where $\mathcal{I}(\cdot)$ is the indicator function, n_j is the total number of cases heard by judge j, and $\hat{\epsilon}$ is the residual from equation (12).

The total sample size for this analysis is 85,865, with 46 judges hearing an average of 1,867 cases each. The judge stringency measure ranges from -0.06 to 0.10; moving from a judge one standard deviation below the mean to a judge one standard above the mean increases conviction probability by 5.2 percentage points. Note that the average conviction rate in the subsample used for the IV analysis is 80.4%. The estimated distribution of residualized judge stringency is given in Appendix B, Figure B.3.

In Table 2, we present summary statistics for the variables used in the IV analysis by conviction status. We would like these statistics to be similar to those for the full sample in order to generalise our findings. However, this subsample contains a higher proportion of misdemeanours and, consequently, higher crime severity for both convicted and acquitted. Nevertheless, it seems reasonable to assume that if convictions for more serious crimes do not impact future criminality, then convictions for lesser crimes would also not affect future criminality.

There is also a discrepancy in race: the proportion of black defendants is lower in the full sample. This difference may raise concerns about extrapolating the IV results of this subsample to the full sample if convictions affect black defendants differently than whites. To address these concerns, the entire IV analysis is also conducted separately for each race group in the Online Appendix. The results of these analyses are consistent with the baseline findings, suggesting that the difference in racial makeup between the two samples is not a concern.

Table 2: Descriptive Statistics (IV Subsample)

	Convicted		Acquitted	
	Mean	St. Dev.	Mean	St. Dev.
Future Criminality	0.81	0.06	0.79	0.07
Previous Criminality	0.81	0.07	0.79	0.07
Neighborhood Effect	0.75	0.02	0.75	0.02
Male	0.64	0.48	0.60	0.49
Black	0.43	0.50	0.43	0.50
Infraction	0.41	0.49	0.38	0.49
Drug Crime	0.09	0.29	0.14	0.34
Property Crime	0.02	0.14	0.04	0.20
Traffic Crime	0.81	0.39	0.64	0.48
Violent Crime	0.07	0.25	0.16	0.36
Court Crime	0.01	0.10	0.02	0.13
Observations	69,012		16,853	
Percentage	80%		20%	

Notes: This table displays means and standard deviations for the sample of defendants from Chesterfield County General District Court (selected according to the criteria given in Section 5.1) used in the IV analysis to test the conditional mean independence assumption.

5.2 IV Assumption Checks

In Table 3, we present the first-stage regression results: a linear probability model of conviction status on judge stringency. Throughout, standard errors are reported in parentheses and are clustered at the judge and defendant level.

In all four regressions, judge stringency has a significant effect. Specifically, being assigned to a judge with a 1 percentage point higher conviction rate increases the chance of conviction by 0.8 percentage points. It is also reassuring to see that this remains stable regardless of the control variables included, providing evidence of the quasi-random assignment of judges.

Table 3: First Stage Regression

	Dependent variable:			
	Convicted			
	(1)	(2)	(3)	(4)
Judge Stringency	0.81***	0.81***	0.80***	0.80***
	(0.14)	(0.15)	(0.15)	(0.15)
Previous Criminality		0.57***	0.20***	0.19***
		(0.04)	(0.03)	(0.03)
White		−0.005	−0.006	0.003
		(0.004)	(0.003)	(0.004)
Male		0.02***	0.02***	0.02***
		(0.003)	(0.003)	(0.003)
Infraction			−0.04***	−0.04**
			(0.01)	(0.01)
Drug Crime			0.05*	0.06*
			(0.02)	(0.02)
Property Crime			−0.03	−0.03
			(0.02)	(0.02)
Traffic Crime			0.17***	0.17***
			(0.02)	(0.02)
Violent Crime			−0.03	−0.03
			(0.02)	(0.02)
ZIP Code Fixed-Effects				✓
Observations	85,865	85,865	85,865	85,865
F-test Statistic	33.2	81.9	62.3	62.4
F-test p-value	0.000	0.000	0.000	0.000

Notes: This table reports first-stage regression results using data from Chesterfield County General District Court. The dependent variable is a binary indicator for conviction. Judge Stringency is the residualized leave-one-out average conviction rate (controlling for year, the crime being a traffic violation, and their interaction). Column (1) regresses conviction status on judge stringency. Column (2) adds defendant characteristics. Column (3) adds the type of crime (court crimes are the reference group) and whether it is an infraction. Column (4) adds ZIP code fixed-effects. Standard errors are in parentheses and are clustered at the judge and defendant level. *, **, and *** indicate 5%, 1%, and 0.1% significance, respectively.

Nevertheless, we formally test this randomisation in Table 4. Here, we regress judge stringency on all case and defendant characteristics and ZIP code fixed-effects. The p-value for the F-statistic of the joint significance of these regressors (including all ZIP code effects) is 0.69, suggesting that judge stringency is unrelated to either case or defendant characteristics. This provides additional support for the validity of the exclusion restriction.

Finally, we check the validity of the monotonicity assumption, which states that defendants who would be convicted by a lenient judge would also be convicted by a more stringent judge, and vice versa for acquittals. In Appendix B, Figure B.4 shows a local linear logit regression of conviction status on judge stringency, demonstrating that the probability of conviction is monotonically increasing and approximately linear.

Table 4: Test of Randomisation

	Dependent variable:
	Judge Stringency
Previous Criminality	0.002 (0.004)
White	−0.0003 (0.0007)
Male	−0.0002 (0.0005)
Infraction	−0.001 (0.001)
Drug Crime	0.003 (0.004)
Property Crime	0.0005 (0.004)
Traffic Crime	0.003 (0.004)
Violent Crime	0.003 (0.004)
ZIP Code Fixed-Effects	✓
Observations	85,865
F-test Statistic	0.70
F-test p-value	0.69

Notes: This table reports results from a test of the randomisation of judge stringency using data from Chesterfield County General District Court. The dependent variable is judge stringency calculated as the residualized leave-one-out average conviction rate of the judge (controlling for year, the crime being a traffic violation, and their interaction). Standard errors are in parentheses and are clustered at the judge and defendant level. *, **, and *** indicate 5%, 1%, and 0.1% significance, respectively.

5.3 IV Results

Table 5: IV Regression

	Dependent variable:
	Future Criminality
Convicted	0.017 (0.026)
Previous Criminality	0.157*** (0.008)
White	0.009*** (0.0008)
Male	0.010*** (0.0007)
Infraction	0.006*** (0.001)
Drug Crime	−0.026*** (0.003)
Property Crime	−0.022*** (0.003)
Traffic Crime	0.032*** (0.006)
Violent Crime	−0.025*** (0.003)
ZIP Code Fixed-Effects	✓
Observations	85,865

Notes: This table reports results from four IV regressions using data from Chesterfield County General District Court. The dependent variable is future criminality calculated using the procedure given in Section 3.2. Conviction is a binary indicator for conviction status and is instrumented by judge stringency. Judge Stringency is the residualized leave-one-out average conviction rate of the judge (controlling for year, the crime being a traffic violation, and their interaction). Standard errors are in parentheses and are clustered at the judge and defendant level. *, **, and *** indicate 5%, 1%, and 0.1% significance, respectively.

Table 5 contains the final IV result. While the statistically insignificant effect of conviction on future criminality is driven partly by the large standard error, the magnitude of the effect is small. Recall that future criminality represents the predicted probability of conviction based solely on measures of future criminal behaviour; thus, the slope coefficient of 0.017 indicates that those who are convicted have future criminality associated with a 1.7 percentage point higher conviction rate. These findings provide evidence that the conditional mean independence assumption is likely to hold, at least approximately.

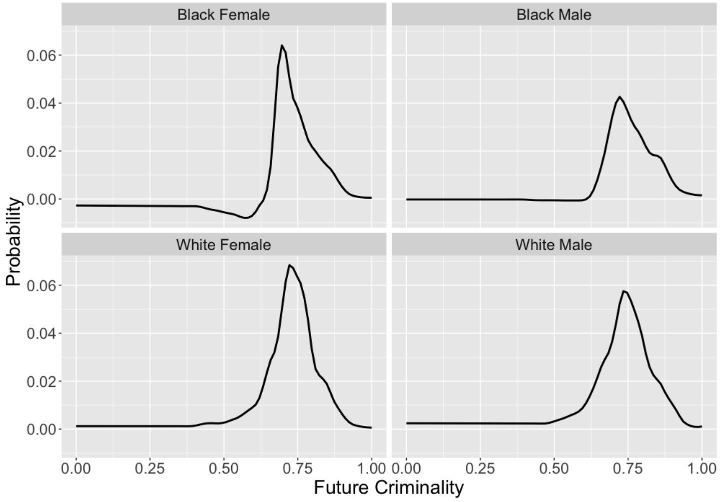
6 Validity of the Large Support Assumption

In this section, we verify the validity of the large support assumption using the check of Section 2.2. We use the same procedure as in Section 4 and the same set of controls, including the race, gender, and previous criminality of the defendant, the neighbourhood effect, whether the crime is an infraction, and the type of crime. We focus on checking whether the large support condition is satisfied for traffic infractions for each race and gender combination where the neighbourhood effect kept at its mean and previous criminality ranges from its 0.1 percentile to its 99.9 percentile.

Figure 6.1 plots $(\partial/\partial v)\, E\left[Y|V=v, X=x\right]$ over the range of the special regressor (future criminality) with previous criminality set at 0.5 (the 0.1 percentile). From these plots, it appears that both tail conditions are satisfied for each race-gender combination, with the upper tail condition only just being satisfied in each case. However, it should be noted that the derivative for black females is slightly negative over a small range of future criminality. This suggests that the relationship between future criminality and the probability of being guilty is not monotonic. While this does not invalidate the identification results, it does cast doubt on the single-index structure (Assumption 3) imposed to test the large support condition. Nonetheless, we could impose a monotonic relationship if we believed it would have a large impact on the results, which does not appear to be the case.

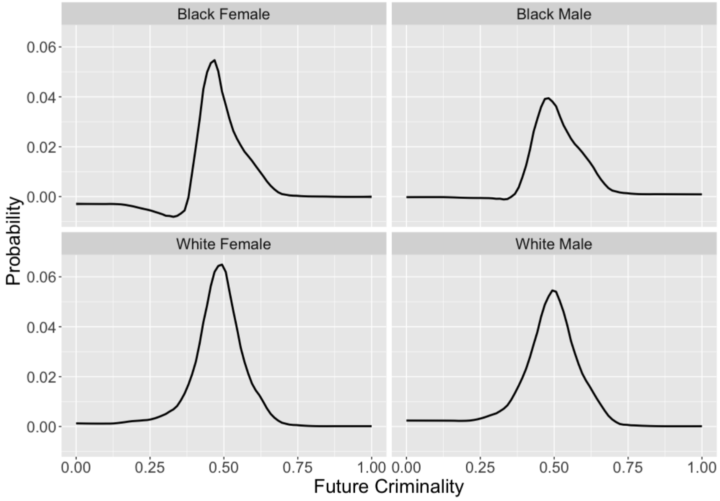
Figure 6.2 displays analogous plots when previous criminality is set to 0.93 (the 99.9 percentile). Here, both tail conditions are comfortably satisfied for each race-gender group. It is intuitive that the mass of the density shifts to the left when previous criminality increases, since the defendants with high future criminality now also have high previous criminality and are thus more likely to be truly guilty of the current crime.

Figure 6.1: Large Support Check (1)



Notes: This figure plots estimates of $\partial E\left[Y|V=v,X=x\right]/\partial v$ over the range of the special regressor (future criminality) with previous criminality set at 0.5 (the 0.1 percentile). The crime type is set to be a traffic infraction, and the neighbourhood effect is set at its mean. A local linear logit estimator on the full dataset (split into the respective race-gender groups) is used.

Figure 6.2: Large Support Check (2)



Notes: This figure plots estimates of $\partial E\left[Y|V=v,X=x\right]/\partial v$ over the range of the special regressor (future criminality) with previous criminality set at 0.93 (the 99.9 percentile). The crime type is set to be a traffic infraction, and the neighbourhood effect is set at its mean. A local linear logit estimator on the full dataset (split into the respective race-gender groups) is used.

7 Conclusion

In this paper, we estimate the likelihood of wrongful conviction and wrongful acquittal using data from over five million court cases in Virginia. Our method is based on reframing the problem in the context of misclassified binary choice models where the misclassification rates are interpreted as type I and type II errors, respectively. We give new nonparametric identification results for these models that admit simple estimators and which are likely to be of independent interest. We also provide a method to test the large support assumption and develop an alternative identification scheme for cases which fail this test. In our empirical context, a thorough discussion of the identification conditions is provided along with evidence of their validity. This includes an analysis of the effect of conviction on future criminality using the leniency of quasi-randomly assigned judges as an IV.

The primary objective of the judicial process is to ensure justice is served by reaching the correct verdict. So it is concerning that the probability of incorrectly convicting a defendant is estimated to be between 63% and 76%, depending on race and gender. However, we must remember that this is an artifact of our sample containing only defendants who have already been arrested and charged with a crime, and not necessarily due to the fallibility of judges. This is also reflected in our wrongful acquittal probability estimate, which ranges from only 4% to 6.5%, depending on race and gender. These findings highlight the complexity and interconnectedness of the various stages of the judicial process and how decisions made at each stage can impact outcomes at later stages. Thus, it is important for policy makers to consider the system as a whole when identifying ways to improve its accuracy.

Regarding future work on this topic, it would be worthwhile to develop a formal test of the large support condition based on our heuristic arguments, as well as investigate whether such a test is applicable in other special regressor models. Inferential procedures for the estimator have also not been developed. If this estimator is to be used in future empirical work, it is crucial to develop methods to gauge the uncertainty of the estimates. Finally, we are silent regarding the performance of judges relative to juries. It would be of great interest to explore if - and when - one type of trial is less prone to error than the other.

References

[1] Acker, J.R. (2017) Taking stock of innocence: Movements, mountains, and wrongful convictions. *Journal of Contemporary Criminal Justice.* 33(1), pp. 8-25.

[2] Altonji, J.G. and R.K. Mansfield (2018) Estimating group effects using averages of observables to control for sorting on unobservables: School and neighbourhood effects. *American Economic Review.* 108(10), pp. 2902-46.

[3] Arnold, D., Dobbie, W. and P. Hull (2021). Measuring racial discrimination in algorithms. *AEA Papers and Proceedings,* 111, pp. 49-54.

[4] Arnold, D., Dobbie, W. and P. Hull (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9), pp. 2992-3038.

[5] Bachoc, F., Preinerstorfer, D. and L. Steinberger (2020) Uniformly valid confidence intervals post-model-selection. *Annals of Statistics.* 48(1), pp. 440-463.

[6] Belloni, A., Chen, D., Chernozhukov, V. and C. Hansen (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica.* 80(6), pp. 2369-2429.

[7] Berry, S.T. and P.A. Haile (2014) Identification in differentiated products markets using market level data. *Econometrica.* 82(5), pp. 1749-1797.

[8] Bhuller, M., Dahl, G.B., Løken, K.V. and M. Mogstad (2020) Incarceration, recidivism, and employment. *Journal of Political Economy.* 128(4), pp. 1269-1324.

[9] Bjerk, D. and E. Helland (2020) What can DNA exonerations tell us about racial differences in wrongful-conviction rates? *Journal of Law and Economics.* 63(2), pp. 341-366.

[10] Canay, I. A., Mogstad, M. and J. Mountjoy (2020) On the use of outcome tests for detecting bias in decision making. *Working Paper.*

[11] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and J. Robins (2018) Double/debiased machine learning for treatment and causal parameters. *Econometrics Journal.* 21(1), pp. 1-68.

[12] Chiricos, T., Barrick, K., Bales, W. and S. Bontrager (2007) The labeling of convicted felons and its consequences for recidivism. *Criminology.* 45(3), pp. 547-581.

[13] Claeskens, G. and I. Van Keilegom (2003) Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics.* 31(6), pp. 1852-1884.

[14] Fan, J., Heckman, N. E. and M.P. Wand (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association.* 90(429), pp. 141-150.

[15] Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics.* pp. 1189-1232.

[16] Frölich, M. (2006) Non-parametric regression for binary dependent variables. *Econometrics Journal.* 9(3), pp. 511-540.

[17] George, T. E. and A. H. Yoon (2017) The gavel gap: Who sits in judgement on state courts. *American Constitution Society for Law and Policy.*

[18] Goh, C. (2018) Rate-optimal estimation of the intercept in a semiparametric sample-selection model. *Working Paper.*

[19] Gross, S.R. and B. O'Brien (2008) Frequency and predictors of false conviction: Why we know so little, and new data on capital cases. *Journal of Empirical Legal Studies.* 5(4), pp. 927-962.

[20] Gross, S.R., O'Brien, B., Hu, C. and E.H. Kennedy (2014) Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences.* 111(20), pp. 7230-7235.

[21] Hansen, B.E. (2022) *Econometrics.* Princeton University Press.

[22] Hartford, J., Lewis, G., Leyton-Brown, K. and M. Taddy (2017) Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning.* 70, pp. 1414-1423.

[23] Hausman, J.A., Abrevaya, J. and F.M. Scott-Morton (1998) Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics.* 87, pp. 239-269.

[24] Heckman, J. J. and S. Navarro (2007) Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics.* 136, pp. 341-396.

[25] Hoekstra, M. and C. Sloan (2022) Does race matter for police use of force? Evidence from 911 calls. *American Economic Review*, 112(3), pp. 827-860.

[26] Hull, P. (2021) What marginal outcome tests can tell us about racially biased decision-making. *Working Paper.*

[27] Khan, S. and D. Nekipelov (2018) Information structure and statistical information in discrete response models. *Quantitative Economics.* 9(2), pp. 995-1017.

[28] Klein, R.W. and R.H. Spady (1993) An efficient semiparametric estimator for binary response models. *Econometrica.* pp. 387-421.

[29] Knox, D., Lowe, W. and J. Mummolo (2020) Administrative records mask racially biased policing. *American Political Science Review.* 114(3), pp. 619-637.

[30] Lewbel, A. (1998) Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica.* pp. 105-121.

[31] Lewbel, A. (2000a) Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics.* 97(1), pp. 145-177.

[32] Lewbel, A. (2000b) Identification of the binary choice model with misclassification. *Econometric Theory.* 16(4), pp. 603-609.

[33] Lewbel, A. and X. Tang (2015) Identification and estimation of games with incomplete information using excluded regressors. *Journal of Econometrics.* 189(1), pp. 229-244.

[34] Magnac, T. and E. Maurin (2007) Identification and information on monotone binary models. *Journal of Econometrics.* 139(1), pp. 76-104.

[35] Mai, C. and R. Subramanian (2017) Price of Prisons: Examining State Spending Trends, 2010-2015. *New York: Vera Institute of Justice.*

[36] Manski, C.F. (1988) Identification of binary response models. *Journal of the American Statistical Association.* 83(403), pp. 729-738.

[37] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and A. Galstyan (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*. 54(6), pp. 1-35.

[38] Mitchell, O., Cochran, J.C., Mears, D.P. and W.D. Bales (2017) Examining prison effects on recidivism: A regression discontinuity approach. *Justice Quarterly*. 34(4), pp. 571-596.

[39] Mueller-Smith, M. (2015) The criminal and labor market impacts of incarceration. *Working Paper*.

[40] Park, K.H. (2017). Do judges have tastes for discrimination? Evidence from criminal courts. *Review of Economics and Statistics*, 99(5), pp. 810-823.

[41] Risinger, D.M. (2006) Innocents convicted: An empirical justified factual wrongful conviction rate. *Journal of Criminal Law and Criminology*. 97, pp. 761.

[42] Spencer, B.D. (2007) Estimating the accuracy of jury verdicts. *Journal of Empirical Legal Studies*. 4(2), pp. 305-329.

[43] Ventura, L.A. and G. Davis (2005) Domestic violence: Court case conviction and recidivism. *Violence Against Women*. 11(2), pp. 255-277.

[44] Zalman, M. (2017) Wrongful Convictions: A Comparative Perspective. *Journal of Contemporary Criminal Justice*. 33(1), pp. 1-7.

Appendix A

Discussion of Mode-Median Coincidence Restriction

In this section, a discussion of the mode-median coincidence restriction of Assumption 4 is provided. We begin with the following theorem:

Theorem 2 For each $x \in supp\,(X)$, let $c_x \in \mathbb{R}$ (for some arbitrary choice c_x) and let $G^*\,(\cdot|x)$ be a function: $[l_V^x, r_V^x] \to [0,1]$. Suppose the following conditions hold: (i) $G^*\,(\cdot|x)$ is non-decreasing and continuously differentiable on $[l_V^x, r_V^x]$;[16] (ii) $\frac{\partial}{\partial v}G^*\,(v|x)$ has a unique maximiser $\bar{v}(x)$ on $[l_V^x, r_V^x]$. Then, for any G^* satisfying (i) and (ii), there exists a pair $(h(x), F_{U|X}(u|x))$ such that a set of random variables (Y^*, X, V, U) satisfies Assumption 2, $\text{Mode}\,[U|X=x]$ satisfies

$$\text{Mode}\,[U|X=x] = c_x, \tag{13}$$

and

$$G^*\,(v|x) = P\,[Y^* = 1|\,(X,V) = (x,v)]\,,$$

for each $v \in [l_V^x, r_V^x]$ and each $x \in supp\,(X)$.

Note that we can set $c_x = 0$ since the choice is arbitrary; however, we consider a non-zero c_x when comparing the conditional mode restriction of equation (13) with Assumption 4 in the main text. Theorem 2 highlights the role of the conditional mode restriction as a location normalisation in monotone discrete choice models to identify $h(x)$ and $F_{U|X}\,(u|x)$. Given the monotonicity of the model, equation (13) does not impose any significant restriction on the functional form of the CPP, $G^*\,(v|x)$, except for the maximiser condition (ii) which is quite mild.

To compare equation (13) and Assumption 4, suppose that Y were observable and thus $G^*\,(v|x)$ is identifiable. Then, letting $c_x = \text{Median}\,[U|X=x]$ gives the restriction in Assumption 4: $\text{Mode}\,[U|X=x] = \text{Median}\,[U|X=x]$. In this case, Assumption 4 would be testable since both $\text{Mode}\,[U|X=x]$ and $\text{Median}\,[U|X=x]$ could be separately identified as $\bar{v}(x)$ and the value v which satisfies $G^*(v|x) = 1/2$, respectively; thus, Assumption 4 could be easily rejected unless $G^*(\bar{v}(x)|x) = 1/2$. However, when Y^* is unobservable - as in our context - $\text{Median}[U|X=x]$ is

[16] We define $\frac{\partial}{\partial v}G^*\,(v|x)$ as the one-sided derivative at each end point of the support.

not identifiable. Therefore, Assumption 4 is not in general testable but imposes a restriction on the form of $G^*(v|x)$, the CPP.

Finally, if $P[Y^* = 1|(X,V) = (x,v)]$ were identifiable, identification of $h(x)$ and $F_{U|X}(u|x)$ could be established, since they are uniquely determined by $G^*(v|x)$ under (i) and (ii), as argued in the proof of Theorem 2.[17] Thus, this theorem can be seen as analogous to Magnac and Maurin's (2007) representation result which is stated under an orthogonality moment condition (corresponding to $E[UX] = 0$ in the present context). We close this section with the theorem's proof:

Proof of Theorem 2 Recall that, given Assumption 3, $P\left[Y^*|\left(X,V\right) = (x,v)\right] = F_{U|X}\left(v + h(x)|x\right)$. Thus, it is sufficient to show that for each $G^*\left(\cdot|x\right)$ which satisfies (i) and (ii), there exists some $(h(x), F_{U|X}(u|x))$ such that $\text{Mode}\left[U|X = x\right] = c_x$ and

$$G^*\left(v|x\right) = F_{U|X}\left(v + h(x)|x\right).$$

Let $\bar{v}(x) := \text{argmax}_{v\in[l_V^x, r_V^x]} \frac{\partial}{\partial v} G^*\left(v|x\right)$ and define $h(x) := c_x - \bar{v}(x)$. Given this $h(x)$, construct $F_{U|X}\left(\cdot|x\right)$ as $F_{U|X}\left(v + h(x)|x\right) := G^*\left(v|x\right)$ for each $v \in [l_V^x, r_V^x]$, or equivalently

$$F_{U|X}\left(u|x\right) := G^*\left(u - h(x)|x\right), \tag{14}$$

for each $u \in [l_V^x + h(x), r_V^x + h(x)]$. By construction, $F_{U|X}\left(u|x\right)$ is differentiable and at $u = c_x$,

$$\frac{\partial}{\partial v} F_{U|X}\left(c_x|x\right) = \frac{\partial}{\partial v} G^*\left(c_x - h(x)|x\right) = \frac{\partial}{\partial v} G^*\left(\bar{v}(x)|x\right).$$

Thus, if $G^*\left(l_V^x|x\right) = 0$ and $G^*\left(r_V^x|x\right) = 1$, we can check that the distribution of $U|X = x$ is fully specified by equation (14) and satisfies equation (13). Otherwise, we can appropriately define the support of $U|X = x$, and the values of $f_{U|X}\left(u|x\right) = \frac{\partial}{\partial u} F_{U|X}\left(u|x\right)$ for $u < l_V^x + h(x)$ or $u > r_V^x + h(x)$, so that $f_{U|X}(u|x) < \frac{\partial}{\partial v} F_{U|X}\left(\bar{v}(x)|x\right)$ for any $u \neq \bar{v}(x)$, $F_{U|X}\left(l_V^x + h(x)|x\right) = G^*\left(l_V^x|x\right)$, and $F_{U|X}\left(r_V^x + h(x)|x\right) = G^*\left(r_V^x|x\right)$. ■

[17] Based on this identification result, a new non/semiparametric estimator for latent-variable binary choice models could be constructed, although this is not pursued in this paper. To the best of our knowledge, there has been no study that considers the conditional mode restriction as in equation (13) for such models.

Derivation of the Integral Form for the Limit Object

Here, we outline how to derive the integral form for the limit object as given in equation (11) in the main text. Note that for each $(x, \tilde{v})$,

$$P\left[Y=1|\left(X,V\right)=\left(x,\tilde{v}\right)\right]=\alpha_1(x)+\left[1-\alpha_1(x)-\alpha_2(x)\right]F_{U|X}\left(\tilde{v}+h(x)|x\right)$$

and

$$\begin{aligned}\int_{\tilde{v}}^{r_V^x}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]dv &= \left[1-\alpha_1(x)-\alpha_2(x)\right]\int_{\tilde{v}}^{r_V^x}f_{U|X}\left(\tilde{v}+h(x)|x\right)dv \\ &= \left[1-\alpha_1(x)-\alpha_2(x)\right]\left[1-F_{U|X}\left(\tilde{v}+h(x)|x\right)\right].\end{aligned}$$

These two equations imply

$$1-\alpha_2(x) = P\left[Y=1|\left(X,V\right)=\left(x,\tilde{v}\right)\right]+\int_{\tilde{v}}^{r_V^x}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]dv.$$

This provides an alternative estimator for $\alpha_2(x)$. However, this equation is based on an arbitrary choice $\tilde{v}$. In the hope of providing a more robust estimation procedure, we take the expectation over $\tilde{v}$. That is,

$$\begin{aligned}&1-\alpha_2(x) \\ =\ & P\left[Y=1|X=x\right]+\int_{l_V^x}^{r_V^x}\left[\int_{\tilde{v}}^{r_V^x}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]dv\right]f_{V|X}\left(\tilde{v}|x\right)d\tilde{v}.\end{aligned}$$

Furthermore, by Fubini's theorem,

$$\begin{aligned}&\int_{l_V^x}^{r_V^x}\left[\int_{\tilde{v}}^{r_V^x}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]dv\right]f_{V|X}\left(\tilde{v}|x\right)d\tilde{v} \\ =\ & \int_{l_V^x}^{r_V^x}\left[\int_{l_V^x}^{v}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]f_{V|X}\left(\tilde{v}|x\right)d\tilde{v}\right]dv \\ =\ & \int_{l_V^x}^{r_V^x}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]F_{V|X}\left(v|x\right)dv.\end{aligned}$$

Thus,

$$1-\alpha_2(x) = P\left[Y=1|X=x\right]+\int_{l_V^x}^{r_V^x}\left(\partial/\partial v\right)P\left[Y=1|\left(X,V\right)=\left(x,v\right)\right]F_{V|X}\left(v|x\right)dv,$$

and the result is obtained. ■

Regression-Based Construction of the Special Regressor

In this section, we discuss the construction of the special regressor V introduced in Section 2.1. As detailed in Section 3.3, we use the future criminal behaviour of the defendant as the special regressor but since there are several measures of future criminality, we construct a scalar V from these measures.

Denote by W a vector of such criminality measures, where W may include discrete components. Recall that one of the basic conditions for V is that it is continuously distributed (supposed in Assumption 1). While V could be simply defined as an average of the components of W, where the averaging may lead to a smoother distribution function (and thus at least approximate continuity), we construct V through a (ML-type) regression of Y on (X, W), where Y is the observable but misclassified version of Y^*, and X is the observable characteristic vector introduced in Section 2.1. Here, we discuss what form the regression should take such that the resulting V is likely to satisfy Assumption 3 or 3′.

First, consider the following nonparametric regression of Y on (X, W):

$$Y = \kappa (X, W) + \epsilon, \tag{15}$$

where $E[\epsilon | X, W] = 0$ and κ is the regression (conditional expectation) function. Note that $\kappa(X, W) \in [0, 1]$ since $Y \in \{0, 1\}$. In our empirical work, V is an estimated probability from a gradient boosted regression tree that lies in $[0, 1]$.

While we could define

$$V_i = \kappa (X_i, W_i), \tag{16}$$

for each individual i, we claim this is not likely to be a sensible choice when W has sufficient predictability for Y. To this end, consider the following slightly strengthened version of Assumption 1:

Assumption 1′

$$E[Y | Y^*, X, W] = E[Y | Y^*, X] \text{ almost surely,}$$

which implies Assumption 1 since V is assumed to be defined as a function of (X, W).

This allows the misclassification error to be written as

$$E\left[Y^{*}-Y\right|(X,W)=(x,w)]=-\alpha_{1}(x)+\left[\alpha_{1}(x)+\alpha_{2}(x)\right]E\left[Y^{*}\right|(X,W)=(x,w)],$$

which can be derived analogously to (1) in Section 2.1. Given (16), the law of iterated expectations leads to

$$E\left[Y^{*}-Y\right|(X,V)=(x,v)]=-\alpha_{1}(x)+\left[\alpha_{1}(x)+\alpha_{2}(x)\right]E\left[Y^{*}\right|(X,V)=(x,v)]. \tag{17}$$

Note, by the definition of κ, we can also write

$$E\left[Y^{*}|X,W\right]=\kappa(X,W)+E\left[Y^{*}-Y|X,W\right]. \tag{18}$$

Taking (18), (17), (16), and the law of iterated expectations, we can write

$$\begin{aligned} E\left[Y^{*}\right|(X,V)=(x,v)] &= v-\alpha_{1}(x)+ \\ &\quad \left[\alpha_{1}(x)+\alpha_{2}(x)\right]E\left[Y^{*}\right|(X,V)=(x,v)]. \end{aligned} \tag{19}$$

Now, suppose there exists some (x,w) such that $\kappa(x,w)=0$, i.e. $Y=0$ can be perfectly predicted. Since $v=\kappa(x,w)$, and $E\left[Y^{*}\right|(X,V)=(x,v)]\geq 0$ we have

$$-\alpha_{1}(x)+\left[\alpha_{1}(x)+\alpha_{2}(x)\right]E\left[Y^{*}\right|(X,V)=(x,0)] \;\geq\; 0.$$

Therefore,

$$E\left[Y^{*}\right|(X,V)=(x,0)] \;\geq\; \frac{\alpha_{1}(x)}{\alpha_{1}(x)+\alpha_{2}(x)}, \tag{20}$$

provided that $0<\alpha_{1}(x)+\alpha_{2}(x)<1$.

On the other hand, suppose there exists some $(x,\tilde{w})$ such that $\kappa(x,\tilde{w})=1$. Then, using (19), we have

$$E\left[Y^{*}\right|(X,V)=(x,1)]=1-\alpha_{1}(x)+\left[\alpha_{1}(x)+\alpha_{2}(x)\right]E\left[Y^{*}\right|(X,V)=(x,1)].$$

Since $E\left[Y^{*}\right|(X,V)=(x,1)]\leq 1$,

$$-\alpha_{1}(x)+\left[\alpha_{1}(x)+\alpha_{2}(x)\right]E\left[Y^{*}\right|(X,V)=(x,1)]\leq 0.$$

Therefore,

$$E\left[Y^{*}\right|(X,V)=(x,1)] \;\; \leq \;\; \frac{\alpha_{1}\left(x\right)}{\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)}. \tag{21}$$

From (20) and (21), $E\left[Y^{*}\right|(X,V)=(x,v)]$ is degenerated to $\alpha_{1}\left(x\right)/\left[\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)\right]$ when it is increasing in v (as supposed in Assumption 3). That is, if predictions from the regression in (15) are used for V, there would be no hope of satisfying the large support condition given Assumptions 1 and 3 and sufficient predictability of Y from W. Thus, it is not sensible to use predictions from a regression of the form in (15) as the special regressor.

Instead, our special regressor is based on a regression of the form

$$Y=\widetilde{\kappa}(X_{1},W)+\tilde{\epsilon},$$

where $E\left[\tilde{\epsilon}|X_{1},W\right]=0$ and X_{1} is a vector consisting of subcomponents in X, and we let $V=\widetilde{\kappa}(X_{1},W)$.

For the above regression, the corresponding expression for (18) is now given by

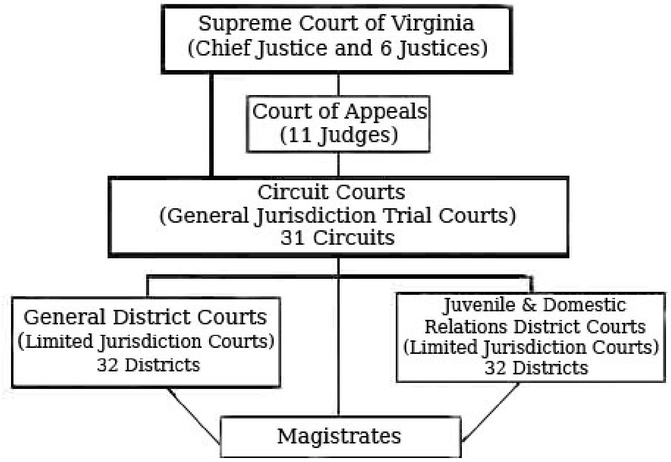
$$E\left[Y^{*}|X_{1},W\right]=\widetilde{\kappa}(X_{1},W)+E\left[Y^{*}-Y|X_{1},W\right].$$

This cannot be (directly) combined with (19) and, consequently, does not result in counterparts to the inequalities in (20) and (21). Thus, the degeneracy result is avoided.

It is worthwhile to note that choosing an X_{1} that is not sufficiently rich is important. This helps to ensure that $E[Y|Y^{*},X_{1},W]\neq E[Y|Y^{*},X_{1}]$, which is required to avoid the degeneracy problem. For this reason, choosing $X_{1}=\emptyset$ (i.e., $V=\widetilde{\kappa}(W)$ without X_{1}) appears to be a sensible choice, and is the choice we make in our empirical setting.

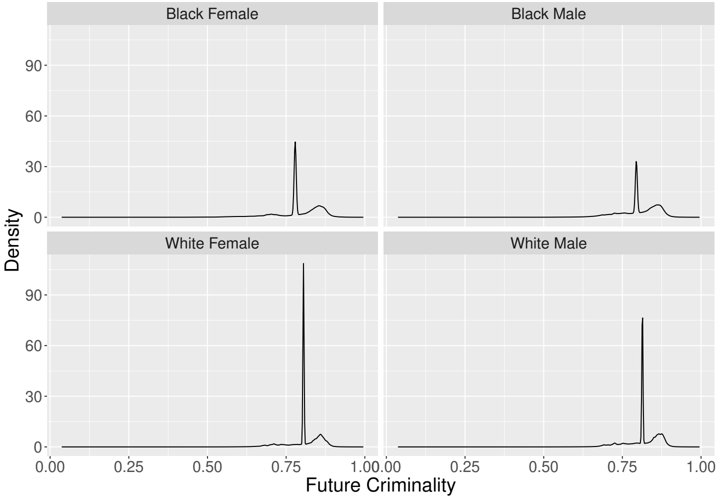
Appendix B

Figure B.1: Virginia Criminal Justice System


Supreme Court of Virginia
(Chief Justice and 6 Justices)
Court of Appeals
(11 Judges)
Circuit Courts
(General Jurisdiction Trial Courts)
31 Circuits
General District Courts
(Limited Jurisdiction Courts)
32 Districts
Juvenile & Domestic
Relations District Courts
(Limited Jurisdiction Courts)
32 Districts
Magistrates

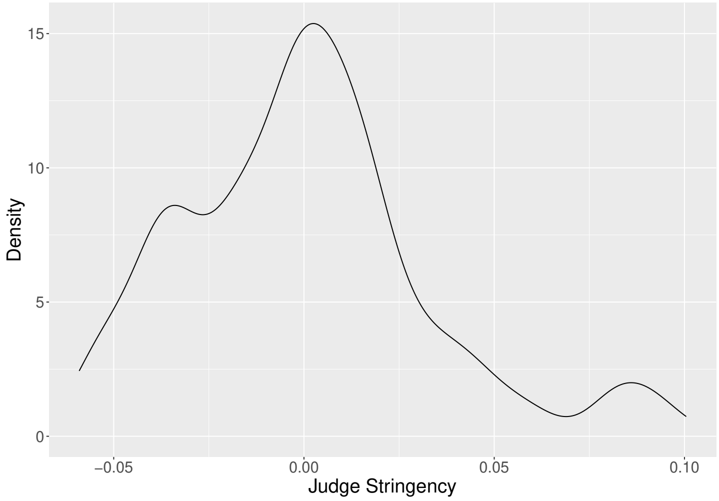
Notes: This figure shows the types of court within the Virginia criminal justice system and how they relate to each other.

Figure B.2: Distribution of Future Criminality



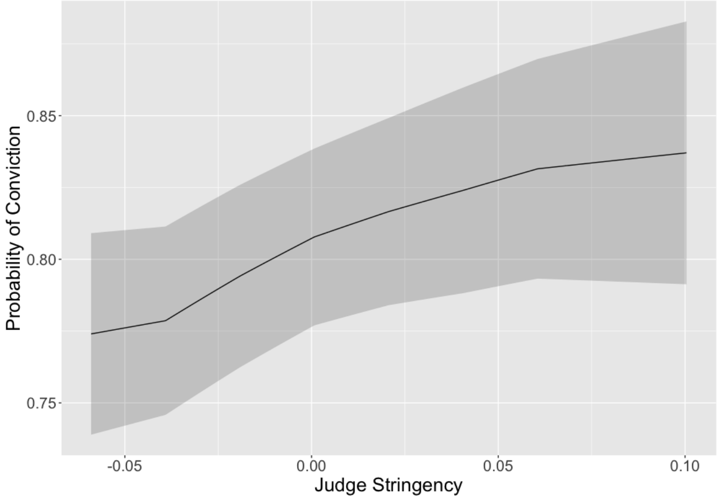
Notes: This figure plots the estimated density of future criminality for each race-gender group constructed using the full sample. The large mass points correspond to 18% of the observations for black females, 12% of observations for black females, 24% of the observations for white females, and 19% for white males.

Figure B.3: Distribution of Judge Stringency



Notes: This figure plots the estimated density of the residualized judge stringency measure constructed using the IV subsample. Judge Stringency is the residualized leave-one-out average conviction rate (controlling for year, the crime being a traffic violation, and their interaction).

Figure B.4: Monotonicity Check



Notes: This figure plots a local linear logit regression of the residualized judge stringency measure on conviction using the IV subsample. The shaded region gives the 95% confidence interval. Judge Stringency is the residualized leave-one-out average conviction rate (controlling for year, the crime being a traffic violation, and their interaction).