Preprint

Estimating Discrimination in Sentencing:
Distinguishing between Good and Bad Controls

Jose Pina-Sánchez, Melissa Hamilton & Peter Tennant

Abstract To minimise confounding bias and disentangle warranted from unwarranted disparities, researchers examining sentencing discrimination have traditionally sought to control for as many legal factors as possible. However, over the past decade, a growing number of scholars have questioned this strategy, noting that many legal factors are themselves subject to judicial discretion and that controlling for them can introduce post-treatment bias. Here, we use directed acyclic graphs (DAGs) to provide a formal and comprehensive assessment of the different types of bias that may arise from different choices of controls. In addition, we propose a new modelling framework to facilitate the selection of controls and reflect the model uncertainty created by the trade-off inherent in judicially-defined legal factors and other factors with a similar dual causal role. We apply this framework examine race disparities in US federal courts and gender disparities in the England and Wales magistrates' court. We find substantial model uncertainty for gender disparities and for race disparities affecting Hispanic offenders, rendering estimates of the latter inconclusive. Disparities against black offenders are more consistent and - under specific conditions - could be interpreted as evidence of direct discrimination.

Keywords Courts · disparities · confounders · mediators · colliders

J. Pina-Sánchez
School of Law, University of Leeds, UK
Liberty Building, University Western Campus, Moorland Rd, Leeds LS3 1DB, UK
ORCID: 0000-0002-9416-6022
E-mail: j.pinasanchez@leeds.ac.uk

M. Hamilton
School of Law, University of Surrey, UK
ORCID: 0000-0002-8593-0017

P. W. G. Tennant
School of Medicine, University of Leeds, UK
ORCID: 0000-0002-1233-8602

1 Introduction

A vast literature documents that offenders from different socio-demographic groups appear to be sentenced differently, suggesting a violation of the principle of equality under the law. Yet, research on sentencing disparities is well known to be fraught with methodological challenges (Baumer, 2013; Wooldredge, 1998). Chief amongst these is confounding bias, which arises from the difficulty ensuring that cases are truly comparable. Otherwise, disparities across socio-demographic groups cannot be confidently interpreted as unwarranted, as they may simply reflect differences in legitimate case characteristics that affect sentence severity (i.e., legal factors). This challenge stems partly from the impossibility to ethically randomise offenders by socio-demographic characteristics in real court settings.[1] Consequently, most research on this area relies on 'selection on observables' designs. The problem is further compounded by two factors: i) even in jurisdictions where judicial discretion is tightly constrained, the set of legal factors that judges may legitimately consider is quite large - potentially infinite if we take the principle of individualisation at heart and accept that 'every case is different' (Cole, 1997) - and ii) the official sentencing data made available to researchers through Sentencing Commissions has often focused on a few characteristics broadly defining the case, thereby omitting many of the legal factors considered by the judge.

The inability to control for all relevant legal factors has been recurrently highlighted as the main methodological limitation in both official government reports on sentencing disparities (Hopkins et al., 2016; Isaac, 2020; United States Sentencing Commission, 2012), and the wider academic literature. As Baumer (2013) notes, after interviewing 25 leading scholars in the field of race disparities in sentencing, four out of five identified 'omitted variable bias' as a problem having a major impact in their findings. Critics have long argued that apparent sentencing disparities across demographic groups are simply the result of group differences in the types of crimes committed (Halevy, 1995; Wilbanks, 1987). Their scepticism is supported by numerous studies showing that estimates of seemingly unwarranted disparities against particular socio-demographic groups are often reduced - or even disappear - once key legal factors are controlled for (Pina-Sánchez et al., 2019; Mitchell, 2005). As a result, until recently the dominant modelling strategy amongst sentencing scholars has been to control for as many legal factors as possible. This advice has been stated explicitly in reviews of the literature (Baumer, 2013; Pina-Sánchez and Linacre, 2016; Zatz, 1987) and has served as an inclusion criteria in more sophisticated meta-analyses (Ferguson and Smith, 2024; Mitchell, 2005).[2]

However, this view should be re-examined. In general, selecting controls simply based on their predictive power is poor modelling practice. At best, it risks overfitting, which reduces statistical power and can introduce multicollinearity. More importantly, ignoring the causal role of variables included in the model is likely to induce bias. A growing number of researchers are increasingly questioning such approach, since many of the legal factors traditionally used as controls in sentencing research are themselves defined at the discretion of sentencers. Consequently, controlling for such factors may inadvertently adjust for the very judicial prejudice under investigation (Kurlychek and Johnson, 2019; Lynch, 2019), resulting in post-treatment bias. Specifically, in studies of race disparities in the U.S. federal courts, Hofer (2019) and Holmes and Feldmeyer (2024) describe a new disciplinary divide. Econometricians tend to favour more parsimonious models and avoid controlling for discretionarily defined legal factors (Fischman and Schanzenbach, 2012; Starr and Rehavi, 2013; Yang, 2015), whereas criminologists typically control for all relevant legal factors. A key point of contention is whether to control for the offence base level or the presumptive sentence. Econometricians argue that the base

[1] A series of studies exploit the random allocation of cases across judges (Abrams et al., 2012; Anderson et al., 1999), a practice designed to prevent 'judge shopping'. Such natural experiments help identify unwarranted disparities across judges. However, to identify sentencing discrimination against offenders from a given socio-demographic group, 'selection on observables' designs remain necessary.

[2] See also Engen and Gainey (2000), who argued that we should also control for interactions between key legal factors increasing model fit.

level provides a more upstream and objective measure of criminality, while criminologists contend that judicial decisions shaping the presumptive sentence - such as reductions for minor participation in the offence or increases for obstruction of justice - are key legal factors that legitimately influence the final decision and should therefore be controlled for (Holmes and Feldmeyer, 2024).

Both arguments are valid but mutually exclusive, leaving sentencing researchers in a 'pick your poison' dilemma whose implications are seldom acknowledged. More troubling still, the challenge is even more complex than previously recognised: many commonly used controls in the sentencing literature can introduce additional forms of bias that have so far gone largely unnoticed. To clarify these mechanisms, we draw on directed acyclic graphs (DAGs), a transparent framework for expressing causal assumptions (Pearl, 2009). DAGs serve two main purposes: i) to represent and communicate causal relationships explicitly, and ii) to establish the role of each variable within the causal process, thereby determining the appropriate set of variables to adjust for when estimating causal effects (Cinelli et al., 2022; VanderWeele and Staudt, 2011). Although now widely used in fields such as public health and epidemiology (Tennant et al., 2021; Textor et al., 2016), DAGs have not yet been systematically applied to sentencing research.[3]

In the next section we provide a succinct introduction to DAGs so that unfamiliar readers can follow the arguments developed in the rest of the article. Specifically, in Section 3 we use DAGs to reconsider the causal role of explanatory variables frequently invoked in sentencing studies, grouping them according to how their inclusion or omission may introduce different types of bias. Building on this classification, in Section 4 we propose a modelling framework to strengthen the robustness of observational studies estimating unwarranted disparities attributable to direct discrimination in sentencing. We then illustrate the framework with two applications: an analysis of racial disparities among drug-trafficking offenders in U.S. federal courts (Section 5.1), and an analysis of gender disparities among shoplifting offenders in the magistrates' courts of England and Wales (Section 5.2).

2 Representing Causal Paths with DAGs

DAGs are non-parametric causal diagrams that express assumed relationships amongst variables within a given context (Fox et al., 2024; Tennant et al., 2021).

Each node represents a variable at a specific point in time, and directed arrows represent hypothesised causal links between them. Paths are formed between nodes – such as between an exposure X and an outcome Y - when they are connected by one or more arrows, and these paths may be open or closed. Open paths transmit statistical associations between variables on that path, closed paths do not.

A causal path is one in which all arrows flow in the same direction. For example, X→M→Y, where X here is the origin node, Y the terminal node, and M a chain node, because it has one incoming and one outgoing arrow. Causal paths are open and transmit causal associations unless an intermediate node is statistically controlled. For instance, controlling for the mediator M in X→M→Y closes the path and removes the causal association between X and Y via M.

A non-causal path is one in which the arrows do not all flow in the same direction. This can occur due to the presence of fork nodes, where two or more arrows diverge from a single node (e.g. C in X←C→Y), or collider nodes, where two or more arrows converge at a single node (e.g. Z in X→Z←Y). Paths that exclusively include chain or fork nodes are open and, when they connect the exposure and outcome they are known as confounding paths. Such paths may be closed by controlling for intermediate nodes, e.g. controlling for confounder C in X←C→Y would close the path and remove confounding bias between X and Y.

In contrast, paths that contain a collider are naturally closed but become open if the collider - or one of its descendants - is controlled for. For example, controlling for collider M in X→M←L→Y opens

[3] See important exceptions in Ward et al. (2016), Pina-Sánchez et al. (2024), or Kemp and Varona (2023).

the path and introduces collider bias between X and Y. Closing this path would require controlling for another intermediate variable, e.g. controlling for mediator-outcome confounder L would close X→M←L→Y.

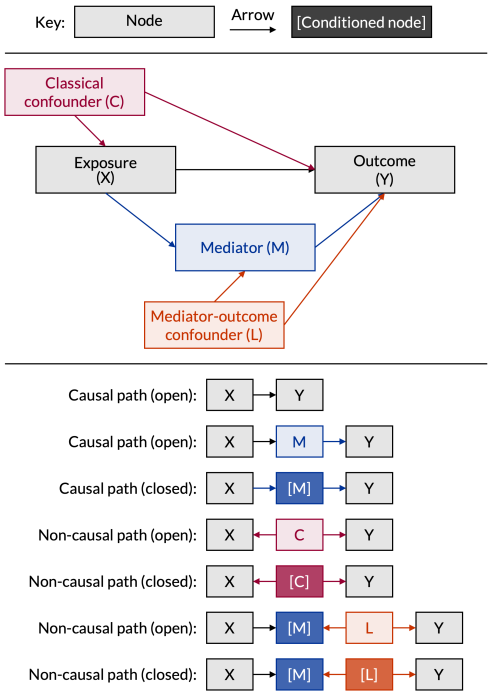

Fig. 1 Representation of common causal and non-causal paths using DAGs

With this knowledge, DAGs can be used to estimate a particular causal effect from an association between two variables after appropriately conditioning on nodes to ensure that all desired causal paths are open and all non-causal paths are closed (Fox et al., 2024; Tennant et al., 2021). Figure 1 illustrates each of the scenarios described above. In the next section, we demonstrate how these structures appear in research on discrimination in sentencing and discuss their implications.

3 How the Choice of Controls Affects Estimates of Discrimination in Sentencing

Researchers examining discrimination in sentencing often fail to explicitly state the causal assumptions underlying their statistical models. Figure 2 depicts the causal model typically implied in studies of this kind (panel A).[4] It shows the central modelling challenge repeatedly highlighted in the literature. Namely, the need to control for all offence and offender characteristics that legally define the case, such as offence seriousness, the offender's criminal history, or the entry of a guilty plea. The usual assumption is that once these case characteristics are controlled for, the estimand of interest - generally, the direct

[4] Figure 2, and all other DAGs presented in this section, are divided in different panels to illustrate the types of bias that arise before (panel A) and after controlling for legal and non-legal factors commonly used in sentencing research (panel B).

effect of the offender's socio-demographic attribute (e.g., race, gender, nationality) on sentence severity - can be identified (panel B).

Without full adjustment, observed differences in sentencing across groups could simply reflect differences in the types of crimes committed. In other words, if legal factors considered by the judge are not fully controlled, the estimated disparities may partly capture 'warranted' rather than 'unwarranted' disparities (Ulmer, 2018). Technically, failing to control for the main case characteristics leaves open the indirect path: *offender's socio-demographics* → *case characteristics* → *sentence*, thereby biasing the effect of interest, *offender's socio-demographics* → *sentence*. Notice how we draw a dashed - rather than a continuous - path from case characteristics to sentence in panel B of Figure 2 to indicate that such path is closed after controlling for case characteristics.

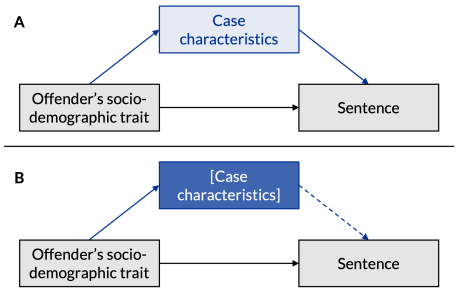

Fig. 2 Approximation of the causal model implicitly invoked in the typical study on sentencing disparities (panel A). To identify the direct effect of the *offender's socio-demographic trait* of interest on *sentence* severity, all relevant *case characteristics* must be controlled (panel B).

In what follows we show how this oversimplified model of the sentencing process obscures important causal mechanisms and, as a result, overlooks additional sources of bias that commonly affect estimates of discrimination in sentencing. To develop this argument, we next expand our DAG to provide a more accurate representation of the sentencing process (Figure 3). We introduce two new nodes, *deliberation* and *judicial prejudice*, drawn as ellipses to indicate that these are unobserved (i.e., latent) variables.

We posit that the causal effect of interest is more accurately conceived as stemming from a hypothetical judicial prejudice that is triggered once the judge forms an initial impression of the offender's socio-demographic group. This hypothetical prejudice then feeds into the process of deliberation, in which all case characteristics are weighed before the sentence is imposed. Because we do not observe judicial prejudice nor the deliberation process, the causal effect of interest is approximated by the estimated effect of the offender's socio-demographic trait on sentence severity, conditional on all other determinant of sentencing disparities that are not attributable to judicial prejudice - represented in this simplified model by the case characteristics.

Figure 3 serves as our reference DAG for illustrating the different causal roles of the main explanatory variables typically included in studies of discrimination in sentencing. We group these variables into two categories of legal factors: pre-defined and judicially-defined case characteristics; and three categories of non-legal factors: non-legal offender characteristics, judge characteristics, and court characteristics.

3.1 Pre-Defined and Judicially-Defined Case Characteristics

An implicit assumption in conventional models of sentencing discrimination is that the case characteristics controlled for in the model represent legal factors exogenous to the judge, who simply weighs

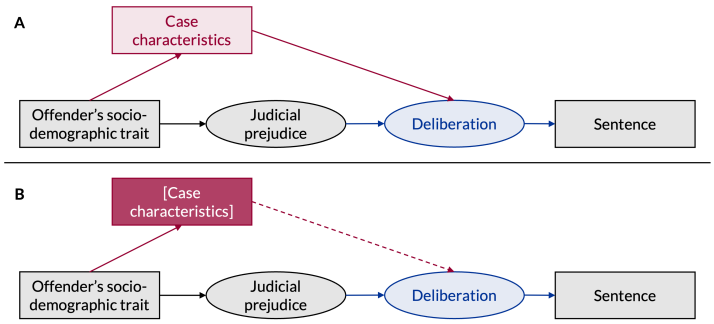

Fig. 3 A more accurate representation of the causal mechanisms underlying sentencing decisions (panel A). The effect of interest runs from *judicial prejudice* to *sentence severity*, mediated by the process of *deliberation*. Because both *judicial prejudice* and *deliberation* are unobserved, this causal effect is approximated by the estimated effect of the *offender's socio-demographic trait* on *sentence* severity. Identifying that effect requires controlling for all relevant *case characteristics* (panel B).

them in determining the appropriate sentence. This simplified view of sentencing - reducing it to an algorithmic processing of inputs into outputs - is not only inaccurate but also highly problematic.

Certain case characteristics can reasonably be viewed as pre-defined or exogenous inputs, at least from the judge's perspective, because they represent objective facts whose presence is not subject to judicial discretion. We refer to these as *pre-defined case characteristics*. Examples include whether the offender entered a guilty plea, whether the offence was committed while on bail, or, in drug cases, the type and quantity of the controlled substance. These factors are typically established through evidence or statutory definitions and would be the same regardless of which judge hears the case.

However, as a growing number of scholars note (Kurlychek and Johnson, 2019; Lynch, 2019; Yang, 2014), many case characteristics commonly treated as legal controls are in fact discretionarily determined by the judge and are therefore endogenous to the very decision-making process under study. For instance, deciding whether an offender's expression of remorse is genuine - and thus warrants a reduced sentence - ultimately rests on the judge's subjective judgment. This assessment is further complicated by the fact that cultural norms shape how remorse is expressed, creating ample room for misinterpretation (Bennett, 2016; Johansen, 2019; Rossmanith, 2015). Consequently, when researchers control for 'remorse', they may not only be adjusting for a legitimately mitigating factor - acknowledgement of harm and a commitment to desist (Bandes, 2016; Maslen, 2015) - but also inadvertently controlling for judicial prejudice. The same concern extends to many other case characteristics whose presence or severity is determined through judicial discretion. Aggravating or mitigating circumstances such as whether a violent offence was premeditated or committed in self-defence, are prime examples. Because these factors are partly products of the judge's deliberation, we refer to them as *judicially-defined case characteristics*.

Figure 4 illustrates this trade-off by presenting a more nuanced depiction of the sentencing process. It distinguishes two key points in judicial deliberation. The first occurs when the judge initially receives the case information (for example, through a pre-sentence report). The second follows the judge's consideration of additional case characteristics that they deem relevant. As discussed, controlling for both pre-defined and judicially-defined case characteristics can help prevent confounding bias, because both sets of factors provide legitimate legal criteria for determining the appropriate sentence. Yet the latter group is determined only after the judge's *initial deliberation* - precisely when any hypothetical judicial prejudice may have already taken effect. Adjusting for these variables can therefore introduce post-treatment bias by blocking part of the causal effect we seek to estimate. This dual causal role is

reflected in Figure 4, where judicially defined case characteristics are shown in purple, combining the red used for confounders and the blue used for mediators.

Specifically, when we control for judicially-defined case characteristics (panel B) we risk blocking the path: *judicial prejudice* → *initial deliberation* → *judicially-defined case characteristics* → *final deliberation* → *sentence*. Conversely, if we do not control for them (panel A) we leave open the backdoor path: *judicial prejudice* ← *socio-demographic trait* → *judicially-defined case characteristics* → *final deliberation* → *sentence*.

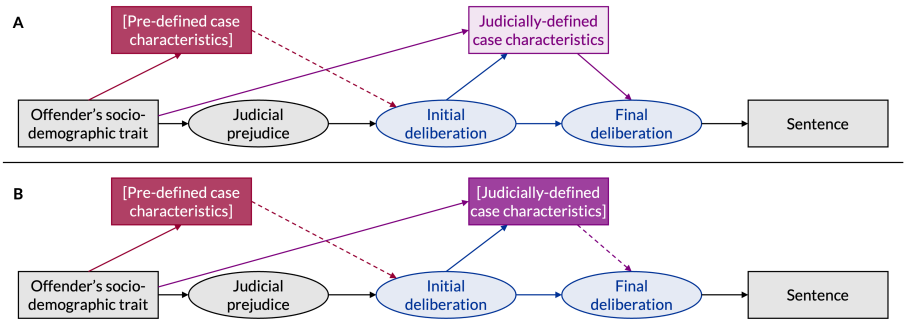

Fig. 4 The trade-off posed by judicially-defined case characteristics (e.g., whether the offender shows remorse or whether the offence was premeditated). Controlling for these factors (panel B) risks post-treatment bias by blocking the path: *judicial prejudice* → *initial deliberation* → *judicially-defined case characteristics* → *final deliberation* → *sentence*. Failing to control for them (panel A), however, risks confounding bias by leaving open the backdoor path: *judicial prejudice* ← *socio-demographic trait* → *judicially-defined case characteristics* → *final deliberation* → *sentence*.

In sum, we should always control for pre-defined case characteristics. By contrast, deciding whether to control for judicially-defined case characteristics entails a trade-off between confounding and post-treatment bias, one that renders estimates of sentencing discrimination ultimately unidentifiable[5]. As we show in Section 4, recognising this limitation does not mean abandoning attempts to estimate discrimination in sentencing. Rather, it requires acknowledging that such estimates are inevitably shaped by model uncertainty. Moreover, although we have presented pre-defined and judicially-defined case characteristics as a simple dichotomy for clarity, in practice that boundary is often blurred and, as discussed in Section 5, highly context-specific.

3.2 Non-Legal Factors

Beyond the offender's socio-demographic trait under analysis and the legal factors discussed above, a wide range of non-legal factors (i.e., variables not specified in the criminal code or invoked in case law) are also commonly used in models of sentencing disparities. We distinguish three broad groups of non-legal factors: i) offender characteristics, ii) judge characteristics, and iii) court and area characteristics.

A vast criminological literature documents substantial differences in crime involvement across socio-demographic groups. Most clearly, men are more likely than women to commit violent crimes (Steffensmeier et al., 2006), and amongst adults, criminal activity declines with age (Steffensmeier et al., 1989). Beyond these well-established patterns, numerous studies have reported differences in criminal involvement across many other socio-demographic dimensions, including race, nationality, employment

[5] If we knew the extent to which these subjectively defined case characteristics are affected by judicial prejudice, we could apply measurement error or misclassification adjustments (Chu et al., 2006; Gustafson, 2003) to overcome this trade-off and still identify the direct effect of judicial prejudice on sentencing.

status, and level of education. At the same time, these same socio-economic traits are also repeatedly found to be associated with sentence severity (Bontrager et al., 2013; Mitchell, 2005; Wermink et al., 2022). Accordingly, when estimating sentencing discrimination against a particular socio-demographic group, researchers should consider the causal role of other offender traits.

Figure 5 extends our standard model by adding a separate node for *other socio-demographic characteristics*. This refers to traits different from the one of interest, which we now call the *focal socio-demographic trait*. For example, when analysing racial disparities, the new node captures all other offender socio-demographic traits aside from race. This additional node plays a similar role to that of the focal trait: it can influence both case characteristics (reflecting the well-documented variations in criminal involvement across socio-demographic groups) and sentence severity through its own form of judicial prejudice (reflecting the sentencing disparities observed across multiple socio-demographic characteristics). In other words, if we ignore relevant socio-demographic characteristics that simultaneously influence both the likelihood of offending and the severity of sentencing, we risk attributing to the focal group trait disparities that actually stem from these other correlated factors.

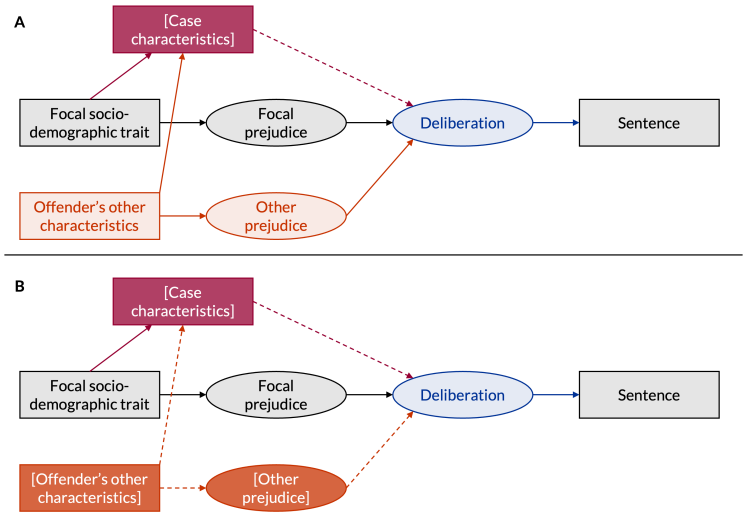

Fig. 5 Other offender socio-demographic characteristics must be controlled for to identify the specific disparity of interest (panel B). This is due to prejudice being unobserved, which conflates different forms of prejudice, but also as result of collider bias from controlling for case characteristics. Specifically, when we control for case characteristics we open up the path: *focal socio-demographic* → *case characteristics* ← *other offender's characteristics* → *other prejudice* → *deliberation* → *sentence*; thereby biasing the effect of interest.

Because both judicial prejudice against the focal socio-demographic trait and other forms of judicial prejudice are unobserved, failing to control for the offender's other socio-demographic characteristics makes it impossible to disentangle their respective effects. Even if focal prejudice could be observed, omitting these additional characteristics would still induce a new form of bias, collider bias. This is a less intuitive but well-documented problem in studies that control for mediators (Richiardi et al., 2013). In the sentencing context, case characteristics act as a mediator of the effect of offender socio-demographics on sentence severity. Once we control for these case characteristics the two nodes representing socio-demographic traits in Figure 5 become associated, and we cannot separate their effects unless we also adjust for the offender's other socio-demographic characteristics (panel B), or for any other variables that jointly affect both case characteristics and sentence severity. Technically, controlling for case characteristics opens the path: *focal socio-demographic* → *case characteristics* ←

other offender's characteristics → *other prejudice* → *deliberation* → *sentence*; biasing the effect of interest.

A different picture emerges when we consider judge characteristics. Although the empirical literature is not always consistent, numerous studies document substantial variation in sentencing across judges (Johnson, 2006; Pina-Sánchez et al., 2019). The evidence base suggests that a judge's gender, race and age can affect sentence severity (Cheng et al., 2023; Johnson, 2014), and that these characteristics may interact with the offender's socio-demographic traits (Cohen and Yang, 2019). Figure 6 illustrates this mechanism, *judge characteristics* can influence the sentence both through the deliberation process (reflecting differences in punitiveness across judges) and through judicial prejudice. Because judicial prejudice itself is unobserved, and we infer its effect on sentence severity examining the observed socio-demographic trait of the offender, controlling for judge characteristics would partially remove the very effect that we aim to estimate (panel B). Accordingly, when seeking to estimate the direct effect of an offender's socio-demographic attribute on sentence severity, judge characteristics should not be controlled for (panel A).

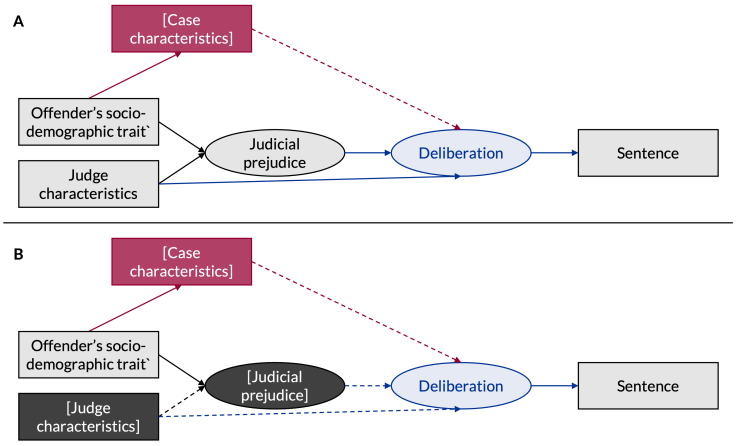

Fig. 6 Judge characteristics influence sentencing both through prejudice and through the deliberation process. Because judicial prejudice is unobserved, controlling for judge characteristics would inadvertently net out differences in prejudice itself (panel B), preventing identification of the effect of interest.

The setting becomes more complex when we consider *court* - or area - *characteristics*. A large body of research shows that factors such as local caseload volume, political voting patterns, or population density can shape both overall punitiveness and the magnitude of ethnic disparities across courts (Ulmer and Kramer, 1996; Ulmer and Johnson, 2004; Wu and Spohn, 2010). These features interact with the criminal justice practitioners that inhabit them to create court cultures (Eisenstein et al., 1988) manifested in different levels of punitiveness and prejudice. However, unlike individual judges - who generally face similar caseloads, especially in jurisdictions with random case assignment - courts located in different areas may confront systematically different offence mixes. Those differences often reflect the underlying patterns of crime in the surrounding community and will be captured in the offenders' case characteristics. Figure 7 captures this additional layer of complexity.

As with other offender socio-demographic characteristics (Figure 5), the effect of courts on the mediator - case characteristics - creates collider bias (panel A) unless court characteristics are controlled for (panel B). This occurs because court characteristics act as a mediator-outcome confounder. Specifically, failing to control for them opens the path: *case characteristics* ← *court characteristics* → *deliberation* → *sentence*. This would justify the inclusion of court or area fixed effects (Rehavi and

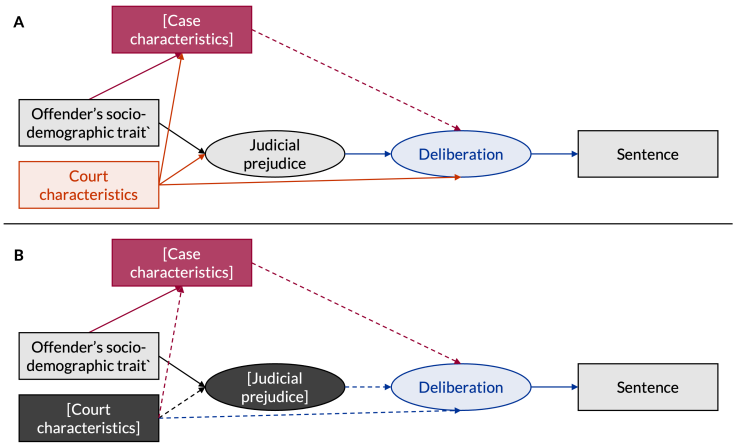

Fig. 7 Trade-off when deciding whether to control for courts or their characteristics. If courts are left uncontrolled (panel A), collider bias arises from controlling for the mediator case characteristics. Specifically, omitting court characteristics opens the path: *case characteristics* ← *court characteristics* → *deliberation* → *sentence*. If court characteristics are controlled (panel B), part of the effect of the offender's socio-demographic trait on sentence severity - operating through judicial prejudice - is inadvertently explained away.

Starr, 2014; Steffensmeier and Demuth, 2006), or alternatively, controlling for court-level characteristics shown to affect sentence severity, such as court caseload or the socio-demographic composition of the area (Caravelis et al., 2011; Ulmer and Johnson, 2004).[6] Yet doing so risks once again absorbing some of the hypothetical judicial prejudice we seek to estimate (panel B), much as when controlling for judge characteristics. Consequently, deciding whether to control for court or area characteristics entails a trade-off closely paralleling that discussed for judicially-defined case characteristics in Section 3.1).

4 Modelling Direct Discrimination in Sentencing

Building on our discussion of how both controlling and failing to control for legal and non-legal factors can introduce bias, this section outlines a modelling framework designed to help researchers minimise such bias, quantify uncertainty, and assess the robustness of their findings. We identify four key steps - not as a prescriptive set of rules, but as a flexible guidelines that can be expanded to incorporate additional sources of model uncertainty, and adapted to different jurisdictions, court types, decision-makers and stages of the sentencing process.

Step 1: Define the Target Estimand

Any causal modelling exercise should begin with a clear definition of the target estimand, i.e the specific effect that we seek to estimate (Lundberg et al., 2021). Yet this is rarely done in the sentencing literature. In practice, the implicit research goal is often to detect discrimination in sentencing, although to avoid criticism researchers commonly frame their objective more modestly, as the estimation of

[6] Another potential source of bias not shown in Figure 7, arises when court characteristics are omitted. The socio-demographic trait of interest is unlikely to be evenly distributed across areas covered by different courts - for example, foreign national and ethnic minority offenders are more likely to reside in urban areas. If court punitiveness is influenced by the urban–rural divide, as seems to be the case (Ruhland and Holmes, 2023), then leaving courts uncontrolled introduces confounding bias.

unwarranted disparities. However, the concept of unwarranted disparities, as generally understood - *"[...] sentencing differences between comparable, similar offenses and offenders that are not attributable to legally prescribed factors"* (Ulmer, 2018, p.2) - lacks the necessary precision. For example, inter-court variability - when certain courts are systematically more punitive than others - constitutes a form of unwarranted disparity (Drápal, 2020; Malin and Tanskanen, 2024), but it does not necessarily imply discrimination against a specific socio-demographic group.

Landmark studies synthesising the evidence on this area of research (Mitchell, 2005; Spohn, 2000; United States Sentencing Commission, 2004) have been clearer in their stated goal by invoking Blumstein et al. (1983) influential distinction between disparities and discrimination: *"Disparity exists when 'like cases' with respect to case attributes - regardless of their legitimacy - are sentenced differently"*, whereas discrimination *"[...] exists when some case attribute that is objectionable (typically on moral or legal grounds) can be shown to be associated with sentence outcomes after all other relevant variables are adequately controlled"* (Blumstein et al., 1983, p.72). This is a useful distinction. However, once again, it is not precise enough. It does not acknowledge the possibility of judicially-defined legal factors (Figure 4), ignores that other offender socio-demographic traits could be biasing the association of interest (Figure 5), and conflates direct discrimination - judicial prejudice - with other forms of structural or indirect discrimination - such as more punitive courts processing a higher number of a particular socio-demographic group of offenders (Figure 7).

Hence, when the research goal is to explore the presence of direct discrimination in sentencing, we suggest adopting the estimand proposed in Section 3, which can be precisely defined as follows: *The average causal effect of* (judicial attitude towards socioeconomic group A) *compared to* (judicial attitude towards socioeconomic group B) *on* (sentence severity) *in* (population P). However, since judicial prejudice - differences in judicial attitudes towards different socio-demographic groups - is unobservable, the above can only be regarded as a theoretical estimand. In practice, we rely on a proxy estimand: *The direct causal effect of* (socioeconomic characteristic A) *compared to* (socioeconomic characteristic B) *on* (sentence severity) *in* (population P), *not mediated through warranted differences in case characteristics.*

Note also that, to define our estimand, we must link it to a clearly specified target population. Researchers often focus not on sentencing practices in a jurisdiction as a whole, but on specific time periods (e.g., before and after the introduction of new guidelines), offence types (e.g., drug offences), or court levels (e.g., magistrates' courts). In such cases, clearly stating the target population is essential to avoid selection bias when generalising findings.

While some subpopulations are easy to conceptualise, others are more ambiguous. For instance, studies often focus on sentence length while excluding the custodial decision itself, potentially introducing selection bias (Bushway et al., 2007; Knox et al., 2020; Pina-Sánchez and Gosling, 2020). Similarly, when sentencers are involved in both conviction and sentencing - as with magistrates in England and Wales - analyses limited to sentencing decisions must clarify that findings do not generalise to the broader decision-making process.

Another common mismatch between the analytical sample and the target population arises in jurisdictions where sentences are negotiated between the defence and prosecution through plea agreements - that is, when sentencing is not solely reserved to judges. As we will discuss in Step 3, clearly identifying the decision-makers under study is essential not only for the sake of transparency, but also because it affects how we classify legal factors as being subjectively defined.

Step 2: Operationalise Sentence Severity

The notion of discrimination in sentencing presupposes the ability to rank sentences according to their relative severity. However, sentencers have access to a wide range of disposal types, many of which are measured on different scales (e.g., fines in monetary units, prison sentences in days), complicating

comparative analysis. This technical challenge has traditionally been addressed in two main ways: by treating sentence severity as a binary variable indicating whether a custodial sentence was imposed, or by analysing the duration of custodial sentences only (Bushway and Piehl, 2001; Ostrom et al., 2008). Both approaches entail a substantial loss of information. The former cannot distinguish severity within each category (custody or non-custody), while focusing on custodial sentences only excludes a significant portion of sentencing decisions, potentially introducing selection bias.

Increasingly, researchers adopt a third strategy: analysing sentence length while coding non-custodial sentences as zeros. This approach allows for the inclusion of the full range of sentencing outcomes, thereby avoiding selection bias, while still enabling differentiation among custodial sentences. Several modelling strategies can be applied to such data, each resting on different assumptions. Given this paper's focus on bias and uncertainty arising from control variable selection, we provide only a brief overview and refer readers to Bushway et al. (2007) for a more detailed discussion.

A key consideration is whether, in the jurisdiction under study, the length of a custodial sentence is determined separately from the disposal type, or whether both are decided simultaneously. In the former case, two-stage models such as Heckman selection models (Heckman, 1976) or hurdle models (Hester and Hartman, 2017) may be appropriate. In the latter, single-stage models such as Tobit regression (King et al., 2010) or negative binomial models (Petersen and Omori, 2020), which treat non-custodial sentences as part of the sentence length distribution, may be more suitable.

An alternative strategy involves estimating the relative severity of each disposal type on a common scale (Irwin-Rogers and Perry, 2015; Leclerc and Tremblay, 2016; Pina-Sánchez and Gosling, 2020). This approach has the unique advantage of distinguishing between non-custodial sentences, thereby maximising the use of available information. Its main drawback lies in the reliance on subjective assessments of severity, which introduces additional assumptions and affects the robustness of findings (Pina-Sánchez and Gosling, 2022).

Step 3: List and Classify the Necessary Controls

Once sentence severity is operationalised and an appropriate outcome model selected, the next step is to identify the explanatory variables required to estimate the target estimand. This involves considering both legal and non-legal factors, and - crucially - distinguishing between variables that should always be controlled for (as doing so can only reduce bias), and those with a dual causal role, which may simultaneously reduce some forms of bias while introducing others.

The list of relevant legal factors can be derived from the criminal code or sentencing guidelines applicable in the jurisdiction under study. To ensure thoroughness, we recommend conducting separate analyses for different offence groups. This is because the relevance of legal factors and the way they are applied vary across offence types, and failing to account for this heterogeneity may lead to model misspecification (Hofer, 2019). For example, aggravating factors such as targeting a vulnerable victim may carry more weight in violent or sexual offences than in property offences, while mitigating factors like the return of stolen goods are specific to the latter. Analysing multiple offence types simultaneously increases sample heterogeneity, which complicates the identification of the causal effect of interest.

Once the relevant legal factors are identified, they should be classified into two categories: pre-defined factors - those presented to the sentencer as part of the case file - and judicially-defined factors - those that may be included or omitted at the sentencer's discretion (Figure 4). As discussed in Section 3.1, this distinction is highly context-dependent. For instance, in England and Wales, pre-trial detention is decided by magistrates. Thus, pre-trial detention (i.e., remand) should be considered a judicially-defined factor when analysing magistrates' courts, but a pre-defined factor when focusing on Crown Court judges.

Similarly, U.S. research on sentencing discrimination involving plea agreements should expand the category of judicially-defined factors to include those shaped by prosecutorial discretion. This logic can

be extended further to include probation and police officers if the research goal is to estimate discrimination across the broader criminal justice system. In essence, the number and nature of discretionarily defined legal factors increase the further upstream the analysis moves in the decision-making process (Kurlychek and Johnson, 2019).[7]

While a comprehensive list of legal factors is essential to detect the presence of unwarranted disparities, it is insufficient for identifying direct discrimination. Researchers must also include offender socio-demographic traits that may be associated with both sentence severity and the trait of interest (Figure 5). Most critically, controls for gender and age should be included, alongside other traits such as citizenship, employment status, and educational attainment.

Additionally, court or area identifiers should be considered to help distinguish direct from indirect discrimination. However, as with judicially-defined factors, controlling for court characteristics may introduce post-treatment bias (Figure 7). Given this similarity, we refer to variables with such dual causal roles - namely, judicially-defined case characteristics and court or area characteristics - as *endogenous controls*, while all other variables in the control set are labelled *exogenous controls*.

Finally, judge characteristics should never be included, as doing so risks controlling away judicial prejudice (Figure 6).[8] More broadly, controls should not be added solely to improve model fit. Any expansion of the control set must be justified in causal terms: variables should be included only if they help reduce a specific form of bias affecting the estimation of direct discrimination in sentencing.

Step 4: Estimate the Model Uncertainty

Because some necessary controls are endogenous to the modelling process - and because controlling for all relevant legal and non-legal factors is ultimately impossible - the target estimand is, strictly speaking, unidentifiable. Nevertheless, we encourage researchers to acknowledge and, where possible, quantify both sources of model uncertainty. Doing so allows us to move beyond reporting point estimates that we know to be biased, and instead offer a bounded range of estimates within which the true effect is likely to lie.

One way to document the uncertainty associated with endogenous controls is to report point estimates for the focal socio-demographic trait across two nested models: one controlling only for the exogenous factors listed in Step 3, and another including both exogenous and endogenous controls. As discussed, both models are misspecified - the former omits relevant variables, while the latter may control for part of the effect we aim to estimate. However, comparing them allows us to assess the uncertainty associated to this trade-off and establish the direction of the bias affecting each of the two modelling strategies.

Since we only suspect - but cannot confirm - that endogenous factors introduce post-treatment bias (e.g., we hypothesise that offender gender may influence whether judges perceive remorse as genuine, but lack direct evidence), a more sophisticated approach may be warranted. Rather than comparing just two models, researchers can use a specification curve analysis (Simonsohn et al., 2020) to estimate a range of models across all - or a random subset of - possible combinations of endogenous controls. This enables us to provide not just an estimate of the bounds within which the true effect might lie,

[7] A similar rationale applies to the data collection process. Legal factors derived from administrative records or case files can generally be considered pre-defined, while those reported directly by sentencers - as in the sentencing surveys conducted by the Sentencing Council for England and Wales - or extracted from sentencing remarks are more likely to be judicially-defined.

[8] Judge characteristics can be informative in studies that seek to explore the mechanisms underlying judicial prejudice, but that is a different research question. When the goal is to estimate the presence of discrimination in sentencing, judge characteristics should not be controlled for.

but by plotting the collection of estimates obtained under different sets of controls as a probability density function we could intuitively assess where the true effect might be more likely to lie.[9]

We also encourage researchers to acknowledge bias arising from failing to control for factors listed in Step 3. Over the last decade, institutions such as the U.S. Sentencing Commission and the Sentencing Council for England and Wales have released increasingly detailed administrative and survey data, alleviating a key limitation affecting past sentencing research (Ulmer, 2011; Zatz, 1987). However, even in these jurisdictions gaps remain, especially regarding non-legal factors. Thus, even studies focusing on relatively homogeneous offence groups and controlling all available case characteristics (Hartley and Miller, 2010; Holland and Prohaska, 2021; Ulmer and Parker, 2020) may still be missing key controls.

Rather than addressing this limitation solely through qualitative caveats - typically placed at the end of the study - we advocate for the use of sensitivity analysis to quantify the potential bias (Lash et al., 2014). As far as we are aware, to date only two studies have applied sensitivity analysis to explore the problem of unobserved confounders in the context of sentencing research (Pina-Sánchez et al., 2024; Ward et al., 2016). Both assume a single unobserved confounder that increases sentence severity (i.e., an aggravating factor), and neither accounts for the variance already explained by observed factors. More sophisticated sensitivity analysis techniques can accommodate multiple unobserved confounders (Groenwold et al., 2016), or even model uncertainty arising from multiple sources simultaneously (Fox et al., 2021; Smith et al., 2021). These approaches could also address additional challenges in sentencing research, such as non-random missing data (Stockton et al., 2024), or misclassification of offender ethnicity.[10]

For researchers unfamiliar with sensitivity analysis, we recommend techniques that are easy to implement across studies. These include, the 'E-value' proposed by VanderWeele and Ding (2017) for effect sizes expressed as odds ratios or risk ratios, and the 'robustness value' introduced by Cinelli and Hazlett (2020) for estimates derived from linear models. Rather than estimating the bias directly, these metrics indicate how strong the association between an unobserved confounder and both the socio-demographic trait of interest and sentence severity would need to be for the observed disparity to disappear entirely.

5 Applications

We now illustrate how the proposed modelling framework can be applied to investigate the presence of direct sentencing discrimination in two distinct contexts: i) against racial minority offenders charged with drug trafficking in U.S. federal courts, and ii) against male offenders charged with shoplifting and sentenced in magistrates' courts in England and Wales.[11] We present two examples, rather than one, to highlight how our proposed modelling framework should be tailored to the specific context under analysis - recognising, in particular, the nuances arising from the jurisdiction, type of court, and decision-makers involved.

[9] Such specification curve analysis could also be extended to account for other sources of model uncertainty, including: i) the choice of functional form used to model sentence severity (Step 2); ii) the specification of more parsimonious models using only a subset of controls from Step 3, which may be preferable in small samples; and iii) more complex models incorporating non-linearities or interaction effects among legal factors (Belton and Dhami, 2023; Engen and Gainey, 2000; Ulmer, 2000) or offender socio-demographic traits (Steffensmeier and Demuth, 2006).

[10] This is a common issue in both U.S. and England and Wales datasets. In the U.S., the reference group of white offenders often includes some Hispanic individuals (Mitchell, 2005), while in England and Wales it may include Gypsy Travellers and other white minorities (Hopkins et al., 2016; Isaac, 2020).

[11] Specifically, the estimand of interest in the first example is the direct causal effect of black and Hispanic drug trafficking offenders, compared to white drug trafficking offenders, on the sentence severity - measured as the length of custodial sentences - imposed by U.S. federal judges and prosecutors offenders from 2019 to 2023. In the second example, the estimand is the direct causal effect of male shoplifting offenders, compared to female shoplifting offenders, on the sentence severity - measured as the probability of receiving a custodial sentence - imposed in the magistrates' courts in England and Wales in 2016.

5.1 Race Disparities in the U.S. Federal Courts

The analysis of race disparities in U.S. federal courts represents one of the most extensively studied topics in sentencing research. Meta-analyses have consistently found statistically significant racial disparities in sentencing outcomes (Ferguson and Smith, 2024; Mitchell, 2005), although the estimated effect sizes tend to be small and highly sensitive to the choice of control variables. In this application, we seek to estimate the presence of direct sentencing discrimination against racial minority offenders convicted of drug trafficking. We focus on drug trafficking offences due to their high volume and relatively limited exposure to difficult-to-measure legal factors, such as offender dangerousness. Importantly, given that only 2.3% of drug cases in U.S. federal courts went to trial in 2023, our definition of direct discrimination encompasses sentencing decisions made by both judges and prosecutors.

The data are sourced from the U.S. Sentencing Commission.[12] We analyse sentences imposed between fiscal years 2020 and 2023, yielding a sample of 72,943 drug trafficking cases. Fourteen variables used in the analysis exhibit item-level missingness, although the extent is minor, with the most affected variable missing only 0.7% of cases. To enable the specification of nested models, we imputed all missing values using multiple imputation via chained equations and predictive mean matching, implemented through the *mice* package in R (Van Buuren, 2018), using the same variables included in the analysis as auxiliary data.[13] Descriptive statistics for these variables - prior to imputation - are presented in Table 1.

Table 1 Descriptive statistics - before imputation - of the variables used in the analysis of race disparities in the U.S. federal courts. The sample is composed of 72,943 drug trafficking offences sentenced from fiscal years 2020 to 2023.

Variable name	Mean[a]	Min.	Max.	Missing	Role
Sentence: prison length	77.1	0	470	0.0%	outcome variable
Offender: white	0.26	0	1	0.2%	focal variable
Offender: black	0.27	0	1	0.2%	focal variable
Offender: Hispanic	0.44	0	1	0.2%	focal variable
Offender: other	0.03	0	1	0.2%	focal variable
Offender: female	0.17	0	1	0.0%	control - exogenous
Offender: age	26.1	18	87	0.0%	control - exogenous
Offender: college education	0.23	0	1	0.7%	control - exogenous
Offender: non-US citizen	0.18	0	1	0.0%	control - exogenous
Drug type: powder cocaine	0.17	0	1	0.2%	control - exogenous
Drug type: crack	0.06	0	1	0.2%	control - exogenous
Drug type: heroin	0.09	0	1	0.2%	control - exogenous
Drug type: marijuana	0.05	0	1	0.2%	control - exogenous
Drug type: meth	0.48	0	1	0.2%	control - exogenous
Drug type: fentanyl	0.10	0	1	0.2%	control - exogenous
Drug type: other	0.05	0	1	0.2%	control - exogenous
Conviction: trial	0.02	0	1	0.0%	control - exogenous
Offence: multiple counts	0.26	0	1	0.1%	control - exogenous
Seriousness: final offence level	26.1	1	6	0.1%	control - endogenous
Criminal history: final category	2.7	1	43	0.0%	control - endogenous
Pre-trial detention	0.76	0	1	1.0%	control - endogenous
Departure: substantial assistance	0.20	0	1	0.0%	control - endogenous
Districts[b]		0	1	0.0%	control - endogenous

[a] The mean of the binary variables represents their proportion in the sample.
[b] The means for each of the 94 districts have been omitted for reasons of space.

[12] Available at: https://www.ussc.gov/research/datafiles/commission-datafiles.
[13] R code and datasets are available at: https://github.com/jmpinasanchez/Sentencing.

Our outcome variable is prison sentence length, measured in months. Probation sentences are coded as 0, and prison sentences are capped at 470 months, consistent with the dataset's coding scheme. The explanatory variables include offence and offender characteristics, as well as district identifiers.[14] We identify five endogenous controls: districts, final offence seriousness level, final criminal history category, pre-trial detention status, and whether a departure was granted for substantial assistance to the prosecution.

Assistance to the prosecution is a relevant mitigating factor but is discretionarily defined by both prosecutors and judges. Pre-trial detention is more complex. In some cases, it reflects factors unrelated to legal severity - such as the offender's ability to post bail.[15] In other cases, it may serve as a proxy for offender dangerousness or risk of reoffending, and thus could be considered a relevant legal factor. However, even under that interpretation, it remains a judicially-defined factor, as pre-trial decisions are typically made by magistrate judges and, when appealed, by district court judges. The final offence seriousness and criminal history variables represent their most complete versions, incorporating discretionary judicial decisions. These include assessments of the offender's role in the offence (e.g., organiser vs. courier), whether the offence targeted a vulnerable victim, and whether prior convictions involved physical force.

We employ negative binomial regression models, which are well-suited to the outcome variable's distribution - right-skewed with a large concentration of zero values. This modelling choice also reflects our assumption that prosecutors and judges jointly determine both the disposal type and its magnitude, rather than through a two-stage process. Table 2 presents results from two nested models: Model 1 includes only exogenous controls, while Model 2 adds the endogenous controls identified above. For simplicity, we focus our analysis on disparities affecting black and Hispanic offenders, whose regression coefficients are highlighted in bold.

Both models indicate small but statistically significant disparities against black offenders. In contrast, estimates for Hispanic offenders vary widely depending on the set of controls included, to the extent that different model specifications yield contradictory conclusions - suggesting either disadvantage or advantage. This highlights a substantial degree of model uncertainty.

Figure 8 illustrates this uncertainty by showing the distribution of estimated disparities for black and Hispanic offenders across 31 unique combinations (excluding the empty set) of the five endogenous controls added to the exogenous control set from Model 1. For black offenders, the 95% interval of the distribution ranges from 1.04 to 1.12, corroborating the small but consistent disadvantage observed in Table 2. For Hispanic offenders, however, the interval spans from 0.82 to 1.18, indicating an extreme degree of model uncertainty and precluding any conclusive interpretation.

To further assess the robustness of the observed disparities against black offenders, we examine whether they could be explained by omitted legal factors, such as drug quantity or whether the offence resulted in death or serious bodily injury. We apply the 'robustness value' technique proposed by (Cinelli and Hazlett, 2020), which is designed for linear models. Accordingly, we replicate Models 1 and 2 using linear regression after log-transforming sentence length.[16] These models yield estimated sentence increases of 15.7% and 6.4% for black offenders, respectively. The corresponding robustness values are 0.041 and 0.023, indicating that an unobserved aggravating factor would need to explain more than 4.1% to 2.3% of the residual variance, and be more than 4.1% to 2.3% more prevalent among black than white offenders, for the observed disparities to be entirely spurious.

Given the relevance of unobserved legal factors such as drug quantity or bodily injury, we expect the first condition to be met. However, the second condition is more questionable. Rehavi and Starr (2014) indicate that in the last three years (2001 to 2003) when quantity of drugs seized at arrest

[14] Most explanatory variables are binary; exceptions include age, final offence level, and criminal history category, which are continuous.

[15] For example, individuals with unlawful immigration status are subject to mandatory pre-trial detention in federal courts. In such cases, controlling for pre-trial detention in a model that already includes offence type may be redundant.

[16] We added one month to all cases prior to log transformation to accommodate probation sentences coded as zero.

Table 2 Results for the negative binomial models on prison sentence length (with probation sentences coded as 0s) for the 72,943 drug trafficking offences sentenced in the US federal courts from fiscal years 2019 to 2023. Effect sizes are reported as incidence rate ratios (IRRs).

	Model 1		Model 2[a]	
	IRR	95% CI	IRR	95% CI
Intercept	46.54	(44.48, 48.69)	4.80	(4.38, 5.26)
Offender: black (ref. white)	**1.09**	(1.07, 1.12)	**1.06**	(1.04, 1.07)
Offender: Hispanic (ref. white)	**0.89**	(0.88, 0.91)	**1.05**	(1.04, 1.06)
Offender: other (ref. white)	0.87	(0.84, 0.91)	1.02	(0.99, 1.06)
Offender: female	0.58	(0.57, 0.59)	0.77	(0.76, 0.79)
Offender: age	1.01	(1.00, 1.01)	1.00	(1.00, 1.00)
Offender: college education	0.87	(0.86, 0.89)	0.96	(0.94, 0.97)
Offender: non-US citizen	1.04	(1.02, 1.06)	0.91	(0.90, 0.93)
Drug type: powder cocaine (ref. crack)	1.16	(1.12, 1.20)	1.05	(1.03, 1.06)
Drug type: heroin (ref. crack)	1.08	(1.04, 1.12)	1.07	(1.04, 1.10)
Drug type: marijuana (ref. crack)	0.53	(0.50, 0.55)	0.97	(0.93, 1.00)
Drug type: meth (ref. crack)	1.73	(1.68, 1.79)	1.07	(1.04, 1.09)
Drug type: fentanyl (ref. crack)	1.12	(1.08, 1.16)	1.11	(1.08, 1.14)
Drug type: other (ref. crack)	0.79	(0.75, 0.82)	0.95	(0.92, 0.98)
Conviction: trial	2.26	(2.15, 2.38)	1.10	(1.07, 1.14)
Offence: multiple counts	1.35	(1.33, 1.38)	1.39	(1.37, 1.40)
Seriousness: final offence level			1.09	(1.09, 1.09)
Criminal history: final category			1.12	(1.12, 1.13)
Pre-trial detention			0.60	(0.59, 0.61)
Departure: substantial assistance			0.67	(0.66, 0.67)

[a] Model 2 includes fixed effects for districts.

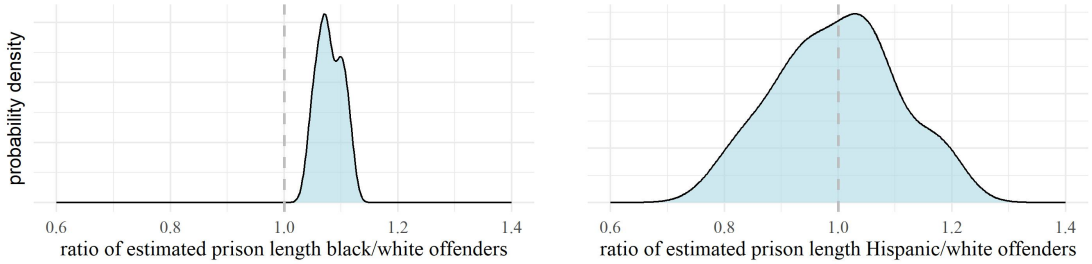

Fig. 8 Model uncertainty associated with the inclusion of endogenous controls in estimating racial disparities. The left panel shows relatively low uncertainty for black offenders, with the entire distribution lying to the right of the vertical dashed line, indicating consistent disadvantage. The right panel shows substantial uncertainty for Hispanic offenders, with the distribution crossing the line, making any conclusion about discrimination highly speculative.

where reliably reported by the Federal Prosecutors Database (EOUSA), no differences were found across racial groups. We are unaware of any studies documenting racial disparities in the prevalence of other aggravating factors such as serious bodily injury. Therefore, under the assumption that black offenders do not systematically traffic larger quantities of drugs or commit offences resulting in bodily injury more frequently than comparable white offenders, our findings may be interpreted as evidence of direct racial discrimination in the sentencing of drug trafficking offenders in U.S. federal courts.

5.2 Gender Disparities in the Magistrates' Courts

Several studies have documented gender disparities in the Crown Court of England and Wales, with male offenders receiving substantially harsher sentences than their female counterparts (Hopkins et al., 2013; Pina-Sánchez and Harris, 2020). However, it remains unclear whether similar disparities exist in the magistrates' courts - the lower-tier courts where 92% of criminal cases are processed (Sturge, 2021). This raises concerns about the external validity of findings based solely on Crown Court data.

In this application, we use the dataset 'Theft from a Shop or Stall', published the Sentencing Council for England and Wales.[17] Specifically, we analyse a sample of 2,116 convicted shoplifting offenders sentenced under the 2016 theft offences sentencing guidelines. The data was collected via self-completed questionnaires[18] administered to magistrates and district judges across 81 magistrates' courts.[19] A key strength of this dataset is that it is completed directly by sentencers, allowing for detailed coverage of the legal factors listed in the sentencing guidelines. However, this also means that the threshold for considering certain legal factors as pre-defined (exogenous) is harder to meet than if they had been reported by administrative staff, such as court clerks.

Our outcome variable is whether the offender received a custodial sentence (either suspended or immediate) or a non-custodial sentence (discharge, fine, or community order), modelled using binary logistic regression. While this is a coarse measure of sentence severity, its use is justified by the low prevalence of immediate custodial sentences for shoplifting offences in magistrates' courts (less than 1% in our sample), and by the absence of sentence length data for suspended custodial sentences. Cases transferred to the Crown Court were excluded, as their final sentencing outcomes are unknown. Similarly, we lack data on conviction decisions, which are also made by magistrates. This mismatch between our target estimand - gender-based direct discrimination in magistrates' court - and our analytical sample introduces potential selection bias. However, we do not expect this bias to be substantial. In 2016 conviction rates for shoplifting offenders were similar across genders (92.2% for female offenders and 92.9% for their male counterparts), and only 0.7% of shoplifting cases were transferred from magistrates' courts to the Crown Court (Ministry of Justice, 2017).

The explanatory variables include offender characteristics and multiple indicators of culpability, harm, aggravating, and mitigating factors present in the case.[20] Although we aim to include all legal factors listed in the sentencing guidelines, some are too rare to meaningfully influence our estimates. Therefore, we exclude case characteristics present in fewer than 1% of cases.[21] Eight variables exhibit item-level missingness, with the highest being guilty plea status (15%). All missing values were imputed using the same multiple imputation approach described in the U.S. federal courts application. Descriptive statistics for the variables - prior to imputation - are reported in Table 3.

In spite of its level of detail, the Sentencing Council datasets have been criticised for omitting difficult-to-measure legal factors, such as offender dangerousness or potential for rehabilitation (Isaac, 2020). In the context of shoplifting offenders, however, dangerousness is unlikely to be a major concern. Moreover, our model includes a proxy for rehabilitation potential - namely, whether the offender has taken steps to address an addiction. One potentially relevant legal factor that appears to be missing in

[17] Available at: https://www.sentencingcouncil.org.uk/research-and-resources/data-collections/magistrates-courts-data-collections/theft-from-a-shop-or-stall/.

[18] https://www.sentencingcouncil.org.uk/wp-content/uploads/Post-guideline-data-collection-form-Theft-from-a-shop-or-stall.pdf.

[19] While the response rate for this specific survey is not reported, previous Sentencing Council surveys have achieved response rates exceeding 60% (Sentencing Council, 2015).

[20] Most variables are binary. Five are ordinal: value of goods (1: 'Up to £10', to 7: '£1001+'), role of offender (1: 'lone' to 4: 'leading'), use of force (1: 'none' to 3: 'high'), level of planning (1: 'none/little' to 3: 'high'), and age (1: '18 to 21' to 5: '50 to 59'). These ordinal variables are treated as continuous in the models for computational simplicity.

[21] We also exclude the overall culpability and harm categories (Step 1 of the sentencing guidelines), as their constituent variables are already included in the model.

Table 3 Descriptive statistics - before imputation - of the variables used in the analysis of gender disparities in the magistrates' courts. The sample is composed of 2,116 shoplifting offenders sentenced in 2016.

Variable name	Mean[a]	Min.	Max.	Missing	Role
Sentence: suspended/immediate custody	0.29	0	1	12.0%	outcome variable
Offender: male	0.72	0	1	0.8%	focal variable
Offender: age band	3.10	1	6	9.0%	control - exogenous
Culpability: level of planning	1.38	1	3	7.1%	control - endogenous
Culpability: use of force	1.06	1	3	13.7%	control - endogenous
Culpability: role of offender	1.37	1	3	7.6%	control - endogenous
Culpability: sophisticated offence	0.02	0	1	0.0%	control - endogenous
Culpability: subject to a banning order	0.01	0	1	0.0%	control - exogenous
Culpability: involvement of others	0.02	0	1	0.0%	control - endogenous
Culpability: mental disorder linked to the offence	0.05	0	1	0.0%	control - endogenous
Harm: value of goods stolen	3.37	1	7	2.5%	control - exogenous
Harm: emotional distress	0.01	0	1	0.0%	control - endogenous
Harm: injury to victim	0.10	0	1	0.0%	control - endogenous
Harm: effect on business	0.38	0	1	0.0%	control - endogenous
Aggravating: previous convictions	0.84	0	1	10.1%	control - endogenous
Aggravating: conceal evidence	0.07	0	1	0.0%	control - endogenous
Aggravating: failure to comply	0.22	0	1	0.0%	control - exogenous
Aggravating: offender on bail	0.17	0	1	0.0%	control - exogenous
Aggravating: additional offences	0.12	0	1	0.0%	control - exogenous
Aggravating: harm to the community	0.20	0	1	0.0%	control - endogenous
Aggravating: professional offending	0.04	0	1	0.0%	control - endogenous
Aggravating: stealing goods to order	0.03	0	1	0.0%	control - exogenous
Mitigating: age / lack of maturity	0.03	0	1	0.0%	control - endogenous
Mitigating: good character	0.02	0	1	0.0%	control - endogenous
Mitigating: no recent convictions	0.08	0	1	0.0%	control - endogenous
Mitigating: financial hardship	0.10	0	1	0.0%	control - endogenous
Mitigating: steps to address addiction	0.13	0	1	0.0%	control - endogenous
Mitigating: mental disorder	0.09	0	1	0.0%	control - endogenous
Mitigating: remorse	0.16	0	1	0.0%	control - endogenous
Mitigating: return of stolen property	0.02	0	1	0.0%	control - exogenous
Mitigating: serious medical condition	0.03	0	1	0.0%	control - endogenous
Mitigating: primary carer	0.02	0	1	0.0%	control - endogenous
Guilty plea: first opportunity	0.96	0	1	15.1%	control - endogenous

[a] Note: The mean of the binary variables represents their proportion in the sample.

assistance to the prosecution.[22] Nevertheless, given that shoplifting offences are typically low in severity and committed by individuals rather than organised groups, we do not expect this to be a common mitigating factor in our sample, nor one that would substantially influence our findings. The dataset is, however, notably sparse in terms of non-legal factors. In particular we lack information on offender socio-demographic characteristics beyond age, and have no data on the court or the geographical location where the sentence was imposed.

To classify controls as exogenous or endogenous, we consider whether each legal factor represents a factual case characteristic or one that could be discretionarily defined by magistrates. Among the variables listed in Table 3, the following are classified as exogenous controls: being subject to a banning order, total value of goods stolen, additional offences admitted, failure to comply with court orders,

[22] Based on informal discussions with members of the Sentencing Council and Crown Court judges, this factor is omitted - despite being explicitly listed in Step 3 of the Sentencing Guidelines - to protect offender anonymity.

offence committed while on bail or licence, stealing goods to order[23], voluntary return of stolen property, serious medical condition, and caring responsibilities.

Some of these - such as failure to comply with a court order or offending while on bail - could arguably be judicially-defined, given that magistrates also adjudicate convictions. However, since prior research indicates no meaningful gender disparities in conviction rates in magistrates' courts, we treat these factors as pre-defined case characteristics. We make an exception for the aggravating factor number of previous convictions. This is because the sentencing guidelines indicate that only previous convictions considered relevant in relation to the offence committed should be counted, we therefore classify the number of previous convictions as judicially-defined, and therefore an endogenous control. All other legal factors listed in Table 3 are also classified as endogenous, as we cannot rule out the possibility that their presence in the case - or their inclusion in the survey form - was influenced by magistrate's perceptions of the offender's gender.

While the magnitude of the estimated gender disparity varies depending on the control set, both models consistently indicate more punitive sentencing for male offenders. Specifically, the odds ratios range from 1.38 to 1.54. When transformed into risk ratios,[24] this corresponds to a range from 1.26 to 1.37, indicating that male offenders are at least 26% more likely to receive a custodial sentence than female offenders.

As in the previous application, a more nuanced exploration of model uncertainty associated with the inclusion of endogenous factors can be obtained through a specification curve. In this case, we identified 20 case characteristics as judicially-defined, resulting in over one million possible model combinations. To simplify the analysis, we re-estimate the model using a random sample of 100 unique combinations. The results are presented in Figure 9, which shows a distribution of odds ratios similar to that observed when comparing Models 1 and 2. Specifically, 95% of the odds ratios fall within the range of 1.34 to 1.58.

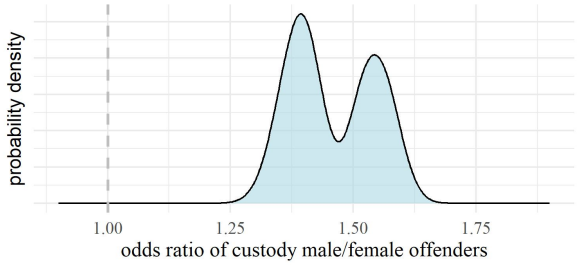

Fig. 9 Model uncertainty associated to the use of judicially-defined case characteristics in the estimation of gender disparities in magistrates' courts. There is a large amount of variability in the distribution of odds ratios making it difficult to identify the specific effect size, however, the distribution remains consistently above 1, indicating a robust disadvantage for male offenders.

An additional benefit of empirically estimating model uncertainty is the ability to assess the influence of specific judicially-defined factors on the estimated disparity. Figure 9 reveals a bimodal distribution, driven primarily by two influential variables. The number of previous convictions substantially reduces the estimated gender disparity by 15.2 percentage points when included in the model, while the presence of injury to the victim increases the estimated disparity by 3.9 percentage points. These findings underscore the importance of carefully considering which judicially-defined factors are included in the model, as they can meaningfully alter the interpretation of sentencing disparities.

[23] Where the offender commits theft with the intention of fulfilling a specific request or demand from another person or group.

[24] Risk ratios were computed using the formula proposed by Zhang and Kai (1998).

Table 4 Results for the logit models specifying the log of the odds of receiving either a suspended or immediate custodial sentence for a sample of 2,116 shoplifting offenders sentenced in the magistrates' courts in 2016. Effect sizes are reported as odds ratios (ORs).

	Model 1		Model 2	
	OR	95% CI	OR	95% CI
Intercept	0.05	(0.03, 0.12)	0.02	(<0.01, 0.05)
Offender: male	**1.54**	(1.20, 1.97)	**1.38**	(1.04, 1.83)
Offender: age band	1.03	(0.93, 1.15)	0.90	(0.80, 1.03)
Culpability: level of planning			1.47	(1.15, 1.87)
Culpability: use of force			1.14	(0.71, 1.82)
Culpability: role of offender			1.06	(0.89, 1.26)
Culpability: sophisticated offence			4.03	(1.71, 9.46)
Culpability: banning order	0.75	(0.26, 2.21)	0.65	(0.18, 2.36)
Culpability: coerced			0.68	(0.24, 1.93)
Culpability: mental disorder			0.45	(0.22, 0.93)
Harm: value of goods stolen	1.38	(1.28, 1.49)	1.31	(1.21, 1.44)
Harm: emotional distress			1.03	(0.39, 2.75)
Harm: injury to victim			6.68	(0.38, 27.8)
Harm: effect on business			0.91	(0.70, 1.18)
Aggravating: previous convictions			1.09	(1.07, 1.11)
Aggravating: conceal evidence			1.14	(0.72, 1.79)
Aggravating: failure to comply	4.99	(3.87, 6.44)	4.92	(3.60, 6.72)
Aggravating: offender on bail	5.33	(4.10, 6.93)	4.36	(3.21, 5.92)
Aggravating: additional offences	1.51	(1.11, 2.05)	1.28	(0.91, 1.80)
Aggravating: harm to the community			4.67	(1.62, 13.46)
Aggravating: professional offending			1.78	(1.01, 3.14)
Aggravating: stealing goods to order	1.80	(1.04, 3.11)	1.11	(0.56, 2.20)
Mitigating: lack of maturity			0.51	(0.21, 1.25)
Mitigating: good character			0.69	(0.19, 2.51)
Mitigating: financial hardship			0.63	(0.39, 1.00)
Mitigating: steps to address addiction			0.46	(0.31, 0.68)
Mitigating: mental disorder			0.66	(0.36, 1.22)
Mitigating: remorse			0.70	(0.48, 1.04)
Mitigating: return of stolen property	0.27	(0.10, 0.74)	0.37	(0.13, 1.09)
Mitigating: serious medical condition			0.89	(0.43, 1.84)
Mitigating: primary carer	0.17	(0.04, 0.74)	0.17	(0.03, 0.91)
Guilty plea: first opportunity	0.74	(0.42, 1.31)	0.81	(0.39, 1.65)

We now turn to explore the robustness of our findings to omitted relevant controls. We apply the 'E-value' technique proposed by (VanderWeele and Ding, 2017), which is designed for discrete models. Specifically, we consider the most conservative of our estimates of gender disparities (Model 2) and acknowledge that our analysis does not control for offender characteristics beyond age - e.g., ethnicity, nationality, employment status, or education level - nor for court identifiers.

The E-value for the lower bound of our estimated risk ratio of receiving a custodial sentence - 1.26 against male offenders - is 1.84. This means that, for our most conservative estimate, for the observed gender disparities to be entirely explained away, an unobserved factor - or set of factors - would need to both increase the probability of receiving a custodial sentence by 84%, and be 84% more prevalent amongst male than female offenders. That represents a substantial threshold, particularly the second condition. We do not expect the socio-demographic profiles of male and female offenders to differ so markedly when conditioning on offence type. For instance, according to Ministry of Justice statistics, the proportion of male theft offenders from ethnic minority backgrounds is only 7% higher than that of

female offenders Ministry of Justice (2017). Similarly, we find it unlikely that male and female offenders are distributed unevenly across courts.

Even if such differences in prevalence could be established, it remains doubtful that the unobserved factors would exert a sufficiently strong influence on sentencing outcomes to meet the required threshold. Existing evidence from England and Wales suggests that race- and deprivation-related disparities in sentencing are statistically significant but relatively modest in magnitude (Lymperopoulou, 2024; Pina-Sánchez et al., 2024).

Therefore, we conclude that - conditional on: i) unobserved legal factors such as dangerousness or assistance to the prosecution not exerting a substantial influence on sentencing decisions for shoplifting offences, and ii) other socio-demographic traits beyond age not being strongly associated with offender gender - there is evidence of direct sentencing discrimination against male offenders in the magistrates' courts. Crucially, however, we cannot generalise this finding to the entire decision-making process in magistrates' courts, as our analysis is limited to sentencing decisions and does not include adjudications of guilt.

6 Discussion

For practical reasons, research on sentencing discrimination has largely relied on observational data, making it particularly sensitive to modelling assumptions. Traditionally, the prevailing approach has been to control for as many legal factors as possible, in order to isolate unwarranted disparities from legitimate sources of variation in sentence severity. More recently, scholars have raised concerns about the potential bias introduced by controlling for legal factors that are discretionarily defined by judges - such as mitigating factors like remorse or aggravating factors like premeditation - arguing that only legal factors exogenous to the judge should be included. The choice between these approaches often reflects disciplinary norms, with criminologists tending to favour the traditional strategy, while econometricians advocate for more parsimonious models (Hofer, 2019).

In this article, we have shown that both approaches aim to reduce important types of bias, but by framing the issue as a binary choice, they fail to address each other's concerns. Ultimately, both are flawed. Using causal diagrams, we illustrate how judicially-defined legal factors often play a dual causal role - acting simultaneously as confounders and mediators of the effect of judicial prejudice on sentence severity. For example, black offenders are less likely to plead guilty than white offenders (Metcalfe and Chiricos, 2018), which correlates with expressions of genuine remorse (Guilfoyle and Pina-Sánchez, 2024). Therefore, failing to control for remorse could lead to confounding bias and wrongly conclude that judges discriminate against black offenders - i.e., such racial disparities might simply be reflecting the sentence reduction warranted from a guilty plea and a genuine expression of remorse. However, the assessment of remorse is itself a subjective decision, potentially influenced by the same judicial prejudice under investigation (Johansen, 2019). In that case, controlling for remorse could also control for judicial prejudice, leading to post-treatment bias and to underestimate the true extent of discrimination. This trade-off renders estimates of sentencing discrimination formally unidentifiable.

By conceptualising the causal mechanisms underlying the sentencing process using DAGs, we also highlighted additional sources of bias stemming from non-legal factors commonly used in sentencing research. Specifically, we show that: i) omitting key offender's socio-demographic characteristics can lead to a less intuitive form of collider bias; ii) controlling for judge characteristics may introduce post-treatment bias; and iii) court or area characteristics can lead to collider bias if unobserved, or post-treatment bias if controlled for.

Beyond documenting these sources of bias, we have used the insights gained to develop a new modelling framework aimed at facilitating more robust and transparent estimation of direct discrimination in sentencing. A key contribution of this framework is its reframing of the trade-off posed by judicially-defined legal factors, and other endogenous non-legal factors like court characteristics.

Rather than choosing between two flawed approaches - controlling for them or excluding them - we advocate for explicitly acknowledging the model uncertainty they introduce. We demonstrate how this can be achieved through specification curve analysis, reporting the distribution of estimates derived from models using different combinations of endogenous controls. By embracing modelling uncertainty, researchers can reduce bias and avoid overconfident interpretations based on naïve point estimates.

To illustrate the framework, we conducted two applications using data from U.S. federal courts and magistrates' courts in England and Wales. These examples underscore the importance of mapping modelling uncertainty, as both the presence and influence of judicially-defined and other endogenous controls vary significantly across contexts. For instance, while estimates of disparities against black offenders in U.S. federal courts appear relatively stable, estimates for Hispanic offenders vary dramatically depending on the control set. Had we followed the criminological approach, we would have concluded that Hispanic offenders receive 5% longer sentences than white offenders; had we followed the econometric approach, we would have concluded they receive 11% shorter sentences. The honest answer is that we do not know which is correct. Without further research into how judicial prejudice is embedded in the construction of legal factors, it is impossible to determine which modelling strategy is superior. Both are biased, and the extent of that bias is unknown. Therefore, without any further information, the best we can do is to report the interval of plausible estimates derived from different model specifications.

Clearly, a more transparent and thoughtful modelling framework for the estimation of discrimination in sentencing is needed. However, it is important to emphasise that the framework presented here should not be viewed as prescriptive set of rules. Rather, it should be understood as a flexible set of guidelines, to be adapted to the specific context under investigations, including: i) the legal framework governing the sentencing process; ii) the socio-demographic realities of the jurisdiction; and iii) the specific decisions and practitioners under analysis. Researchers must exercise discretion in determining which factors are relevant, the causal role they may play, the appropriateness of controlling for them, and the implications of failing to do so.

Similarly, the modelling framework proposed is not definitive, just a step forward in the path towards improving the transparency and robustness of research in sentencing disparities. While our framework focuses on the choice of controls, this is only one source of model uncertainty in sentencing research. Additional steps could be incorporated to: i) address missing data - still a widespread and largely overlooked issue in sentencing research (Stockton et al., 2024); ii) implement pre-registration protocols to reduce researcher bias (Baldwin et al., 2022) - we are only aware of three sentencing studies that have employed pre-registered analytical strategies (Ferguson and Smith, 2024; Pina-Sánchez et al., 2024; Pina-Sánchez and Brunton-Smith, 2025); iii) present effect sizes more intuitively - for example, replacing odds ratios with marginal effects expressed as probabilities (Mize et al., 2019); and iv) expand specification curve analyses to account for model uncertainty stemming from different, yet equally defensible, functional forms used to model the outcome variable.

To conclude, even under the most rigorous modelling strategy and with the most detailed dataset, the precise level of discrimination in sentencing can never be fully identified. Yet, this epistemic limitation should not lead to scientific nihilism or discourage efforts to investigate the issue. Discrimination in sentencing undermines the principle of equal treatment, which lies at the heart of justice. Moreover, perceptions of discrimination - whether grounded on reality or not - can have real-word consequences, eroding trust in public institutions and alienating segments of the population.

Thus, we must reject both the overconfidence that treats single-point estimates as definitive, and the constructivist perspective that dismisses any attempt to quantify judicial discrimination. Instead, we advocate for a middle path: abandoning the pursuit of a singular estimate of judicial prejudice and focusing instead on documenting the uncertainty surrounding the estimation process. This approach offers a more realistic understanding of the problem, and - when applied carefully, and when the resulting evidence is sufficiently clear - can allow researchers to make confident claims about the presence or absence of discrimination in sentencing.

References

Abrams DS, Bertrand M, Mullainathan S (2012) Do judges vary in their treatment of race? The Journal of Legal Studies 41(2):347–383

Anderson JM, Kling JR, Stith K (1999) Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. The Journal of Law and Economics 42(1):271–308

Baldwin JR, Pingault JB, Schoeler T, Sallis HM, Munafò MR (2022) Protecting against researcher bias in secondary data analysis: Challenges and potential solutions. European Journal of Epidemiology 37(1):1–10

Bandes SA (2016) Remorse and criminal justice. Emotion Review 8(1):14–19

Baumer EP (2013) Reassessing and redirecting research on race and sentencing. Justice Quarterly 30(2):231–261

Belton I, Dhami M (2023) The role of character-based personal mitigation in sentencing judgments. Journal of Empirical Legal Studies

Bennett CD (2016) The role of remorse in criminal justice. In: Tonry M (ed) Oxford Handbook Online in Criminology and Criminal Justice

Blumstein A, Cohen J, Martin SE, Tonry MH (1983) Research on sentencing: The search for reform, vol 1. National Academy Press, Washington, D.C.

Bontrager S, Barrick K, Stupi E (2013) Gender and sentencing: A meta-analysis of contemporary research. Journal of Gender, Race & Justice 16:349

Bushway SD, Piehl AM (2001) Judging judicial discretion: Legal factors and racial discrimination in sentencing. Law & Society Review 35(4):733–764

Bushway SD, Johnson BD, Slocum LA (2007) Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. Journal of Quantitative Criminology 23(2):151–178

Caravelis C, Chiricos T, Bales W (2011) Static and dynamic indicators of minority threat in sentencing outcomes: A multi-level analysis. Journal of Quantitative Criminology 27:405–425

Cheng KKY, Ri S, Chengchen H (2023) Judges' characteristics and sentencing in Hong Kong. Criminology & Criminal Justice

Chu H, Wang Z, Cole SR, Greenland S (2006) Sensitivity analysis of misclassification: A graphical and a Bayesian approach. Annals of Epidemiology 16(11):834–841

Cinelli C, Hazlett C (2020) Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B 82(1):39–67

Cinelli C, Forney A, Pearl J (2022) A crash course in good and bad controls. Sociological Methods & Research

Cohen A, Yang CS (2019) Judicial politics and sentencing decisions. American Economic Journal: Economic Policy 11(1):160–191

Cole K (1997) The empty idea of sentencing disparity. Northwestern University Law Review 91(4):1336–1341

Drápal J (2020) Sentencing disparities in the Czech Republic: Empirical evidence from post-communist Europe. European Journal of Criminology 2(17):151–174

Eisenstein J, Flemming RB, Nardulli PF (1988) The contours of justice: Communities and their courts. University Press of America, Boston, MA: Little, Brown.

Engen RL, Gainey RR (2000) Modelling the effects of legally relevant and extralegal factors under sentencing guidelines: The rules have changed. Criminology 38(4):1207–1230

Ferguson CJ, Smith S (2024) Race, class, and criminal adjudication: Is the US criminal justice system as biased as is often assumed? A meta-analytic review. Aggression and Violent Behavior

Fischman JB, Schanzenbach MM (2012) Racial disparities under the federal sentencing guidelines: The role of judicial discretion and mandatory minimums. Journal of Empirical Legal Studies 9(4):729–764

Fox M, Arah OA, Swanson S, Viallon V (2024) Causal diagrams to evaluate sources of bias. In: Statistical Methods in Cancer Research Volume V: Bias Assessment in Case–Control and Cohort

Studies for Hazard Identification, International Agency for Research on Cancer
Fox MP, MacLehose RF, Lash TL (2021) Applying quantitative bias analysis to epidemiologic data. Springer, New York
Groenwold RH, Sterne JA, Lawlor DA, Moons KG, Hoes AW, Tilling K (2016) Sensitivity analysis for the effects of multiple unmeasured confounders. Annals of Epidemiology 26(9):605–611
Guilfoyle E, Pina-Sánchez J (2024) Racially determined case characteristics: Exploring disparities in the use of sentencing factors in England and Wales. The British Journal of Criminology
Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments. CRC Press
Halevy T (1995) Racial discrimination in sentencing: a study with dubious conclusions. Criminal Law Review pp 267–271
Hartley RD, Miller M (2010) Crack-ing the media myth: Reconsidering sentencing severity for cocaine offenders by drug type. Criminal Justice Review 35(1):67–89
Heckman J (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5:475–492
Hester R, Hartman T (2017) Conditional race disparities in criminal sentencing: A test of the liberation hypothesis from a non-guidelines state. Journal of Quantitative Criminology 33:77–100
Hofer PJ (2019) Federal sentencing after Booker. Crime and Justice 48(1):137–186
Holland MM, Prohaska A (2021) Gender effects across place: A multilevel investigation of gender, race/ethnicity, and region in sentencing. Race and Justice 11(1):91–112
Holmes B, Feldmeyer B (2024) Modeling matters: Comparing the presumptive sentence versus base offense level approaches for estimating racial/ethnic effects on federal sentencing. Journal of Quantitative Criminology 40(2):395–420
Hopkins K, Light M, Lvbakke J (2013) Analysis of gender as a factor associated with custodial sentences for breach of a court order. Ministry of Justice URL https://www.gov.uk/government/statistics/women-and-the-criminal-justice-system-2013
Hopkins K, Uhrig N, Colahan M (2016) Associations between ethnic background and being sentenced to prison in the crown court in England and Wales in 2015. Tech. rep., Ministry of Justice, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/639261/bame-disproportionality-in-the-cjs.pdf
Irwin-Rogers K, Perry T (2015) Exploring the impact of sentencing factors on sentencing domestic burglary. In: Roberts J (ed) Sentencing Guidelines: Exploring Sentencing Practice in England and Wales, Palgrave, Basingstoke, pp 213–239.
Isaac A (2020) Investigating the association between an offender's sex and ethnicity and the sentence imposed at the Crown Court for drug offences. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/wp-content/uploads/Sex-and-ethnicity-analysis-final-1.pdf
Johansen LV (2019) 'Impressed' by feelings: How judges perceive defendants' emotional expressions in danish courtrooms. Social & Legal Studies 28(2):250–269
Johnson BD (2006) The multilevel context of criminal sentencing: Integrating judge-and county-level influences. Criminology 44(2):259–298
Johnson BD (2014) Judges on trial: A reexamination of judicial race and gender effects across modes of conviction. Criminal Justice Policy Review 25(2):159–184
Kemp S, Varona D (2023) Foreign and dangerous? Unpackingthe role of judges and prosecutors insentencing disparities in Spain. The British Journal of Criminology 63:984–1002
King R, Johnson K, McGeever K (2010) Demography of the legal profession and racial disparities insentencing. Law Society Review 44(1):1–32
Knox D, Lowe W, Mummolo J (2020) Administrative records mask racially biased policing. American Political Science Review 114(3):619–637

Kurlychek MC, Johnson BD (2019) Cumulative disadvantage in the American criminal justice system. Annual Review of Criminology 2:291–319
Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S (2014) Good practices for quantitative bias analysis. International Journal of Epidemiology 43(6):1969–1985
Leclerc C, Tremblay P (2016) Looking at penalty scales: how judicial actors and the general public judgepenal severity. Canadian Journal of Criminology and Criminal Justice 58(3):354–384
Lundberg I, Johnson R, Stewart BM (2021) What is your estimand? Defining the target quantity connects statistical evidence to theory. American Sociological Review 86(3):532–565
Lymperopoulou K (2024) Ethnic inequalities in sentencing: Evidence from the crown court in England and Wales. The British Journal of Criminology
Lynch M (2019) Focally concerned about focal concerns: A conceptual and methodological critique of sentencing disparities research. Justice Quarterly 36(7):1148–1175
Malin T, Tanskanen M (2024) Exploring sentencing disparities in the Nordic context: A multilevel analysis of court-and judge-level variation in sentences of 'aggravated driving under the influence' in Finnish district courts. Criminology & Criminal Justice
Maslen H (2015) Remorse, penal theory and sentencing. Bloomsbury Publishing.
Metcalfe C, Chiricos T (2018) Race, plea, and charge reduction: An assessment of racial disparities in the plea process. Justice Quarterly 35(2):223–253
Ministry of Justice (2017) Criminal justice system statistics quarterly: December 2016
Mitchell O (2005) A meta-analysis of race and sentencing research: Explaining the inconsistencies. Journal of Quantitative Criminology 21(4):439–466
Mize TD, Doan L, Scott Long J (2019) A general framework for comparing predictors and marginal effects across models. Sociological Methodology 49(1):152–189
Ostrom B, Ostrom C, Hanson R, Kleiman M (2008) Assessing Consistency and Fairness in Sentencing: A Comparative Study in Three States. National Institute of Justice, Washington
Pearl J (2009) Causality. Cambridge University Press
Petersen N, Omori M (2020) Is the process the only punishment?: Racial–ethnic disparities in lower-level courts. Law & Policy 42(1):56–77
Pina-Sánchez J, Gosling JP (2020) Tackling selection bias in sentencing data analysis: A new approach based on a scale of severity. Quality & Quantity 54:1047–1073
Pina-Sánchez J, Harris L (2020) Sentencing gender? Investigating the extent and origin of sentencing gender disparities in the Crown Court. Criminal Law Review 1:3–28
Pina-Sánchez J, Linacre R (2016) Refining the measurement of consistency in sentencing: A methodological review. International Journal of Law, Crime and Justice 44:68–87
Pina-Sánchez J, Grech D, Brunton-Smith I, Sferopoulos D (2019) Exploring the origin of sentencing disparities in the Crown Court: Using text mining techniques to differentiate between court and judge disparities. Social Science Research 84:102343
Pina-Sánchez J, V RJ, Sferopoulos D (2019) Does the Crown Court discriminate against Muslim-named offenders? A novel investigation based on text mining techniques. British Journal of Criminology 59(3):718–736
Pina-Sánchez J, Geneletti S, Veiga A, Morales A, Guilfoyle E (2024) Can ethnic disparities in sentencing be taken as evidence of judicial discrimination? Journal of Legal Research Methodology 3(1):54–82
Pina-Sánchez J, Brunton-Smith I (2025) What is the external validity of sentencing research? a multi-level meta-analysis of race and gender disparities. SocArXiv URL https://doi.org/10.31235/osf.io/5d2bq_v2
Pina-Sánchez J, Gosling JP (2022) Enhancing the measurement of sentence severity through expert knowledge elicitation. Journal of Legal Research Methodology 2(1):26–45
Pina-Sánchez J, Morales A, Guilfoyle E, Veiga A, Geneletti S (2024) The interrelationship between area deprivation and ethnic disparities in sentencing. Analyses of Social Issues and Public Policy

Rehavi MM, Starr SB (2014) Racial disparity in federal criminal sentences. Journal of Political Economy 122(6):1320–1354
Richiardi L, Bellocco R, Zugna D (2013) Mediation analysis in epidemiology: Methods, interpretation and bias. International Journal of Epidemiology 42(5):1511–1519
Rossmanith K (2015) Affect and the judicial assessment of offenders: Feeling and judging remorse. Body & Society 21(2):167–193
Ruhland E, Holmes B (2023) An examination of sentencing outcomes in rural and urban locations. American Journal of Criminal Justice 48:701–722
Sentencing Council (2015) Annex b: Quality and methodology note. Tech. rep., URL https://www.sentencingcouncil.org.uk/wp-content/uploads/CCSS-Annex-B1.pdf
Simonsohn U, Simmons JP, Nelson LD (2020) Specification curve analysis. Nature Human Behaviour 4(11):1208–1214
Smith LH, Mathur MB, VanderWeele TJ (2021) Multiple-bias sensitivity analysis using bounds. Epidemiology 32(5):625–634
Spohn C (2000) Thirty years of sentencing reform: The quest for a racially neutral sentencing process. Criminal justice 3(1):427–501
Starr SB, Rehavi MM (2013) Mandatory sentencing and racial disparity: Assessing the role of prosecutors and the effects of booker. Yale Law Journal 123(1):2–80
Steffensmeier D, Demuth S (2006) Does gender modify the effects of race–ethnicity on criminal sanctioning? Sentences for male and female white, black, and Hispanic defendants. Journal of Quantitative Criminology 22:241–261
Steffensmeier D, Zhong H, Ackerman J, Schwartz J, Agha S (2006) Gender gap trends for violent crimes, 1980 to 2003: A UCR-NCVS comparison. Feminist Criminology 1(1):72–98
Steffensmeier DJ, Allan EA, Harer MD, Streifel C (1989) Age and the distribution of crime. American Journal of Sociology 94(4):803–831
Stockton B, Strange CC, Harel O (2024) Now you see it, now you don't: A simulation and illustration of the importance of treating incomplete data in estimating race effects in sentencing. Journal of Quantitative Criminology 40:563–590
Sturge G (2021) Court statistics for England and Wales. Tech. rep., URL https://researchbriefings.files.parliament.uk/documents/CBP-8372/CBP-8372.pdf
Tennant PW, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, Harrison WJ, Keeble C, Ranker LR, Textor J, Tomova GD, Gilthorpe MS, Ellison GT (2021) Use of directed acyclic graphs (dags) to identify confounders in applied health research: Review and recommendations. International Journal of Epidemiology 50(2):620–632
Textor J, Van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH (2016) Robust causal inference using directed acyclic graphs: The R package 'dagitty'. International Journal of Epidemiology 45(6):1887–1894
Ulmer J (2018) Race, ethnicity, and sentencing. Oxford Research Encyclopedia of Criminology
Ulmer JT (2000) The rules have changed-so proceed with caution: A comment on Engen and Gainey's method for modeling sentencing outcomes under guidelines. Criminology 38(4):1231–1243
Ulmer JT (2011) Recent developments and new directions in sentencing research. Justice Quarterly 29(1):1–40
Ulmer JT, Johnson B (2004) Sentencing in context: A multilevel analysis. Criminology 42:137–177
Ulmer JT, Kramer JH (1996) Court communities under sentencing guidelines: Dilemmas of formal rationality and sentencing disparitiy. Criminology 34(3):289–464
Ulmer JT, Parker BR (2020) Federal sentencing of Hispanic defendants in changing immigrant destinations. Justice Quarterly 37(3):541–570
United States Sentencing Commission (2004) Fifteen years of guidelines sentencing URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-projects-and-surveys/miscellaneous/15-year-study/15_year_study_full.pdf

United States Sentencing Commission (2012) Report to the congress: Continuing impact of United States v. Booker on federal sentencing. URL https://www.ussc.gov/research/congressional-reports/2012-report-congress-continuing-impact-united-states-v-booker-federal-sentencing

Van Buuren S (2018) Flexible imputation of missing data. CRC press

VanderWeele TJ, Ding P (2017) Sensitivity analysis in observational research: Introducing the e-value. Annals of Internal Medicine 167(4):268–274

VanderWeele TJ, Staudt N (2011) Causal diagrams for empirical legal research: A methodology for identifying causation, avoiding bias and interpreting results. Law, Probability and Risk 10(4):329–354

Ward JT, Hartley RD, Tillyer R (2016) Unpacking gender and racial/ethnic biases in the federal sentencing of drug offenders: A causal mediation approach. Journal of Criminal Justice 46:196–206

Wermink H, Light MT, Krubnik AP (2022) Pretrial detention and incarceration decisions for foreign nationals: A mixed-methods approach. European Journal on Criminal Policy and Research 28(3):367–380

Wilbanks W (1987) The Myth of a Racist Criminal Justice System. Brooks/Cole, Monterey

Wooldredge JD (1998) Analytical rigor in studies of disparities in criminal case processing. Journal of Quantitative Criminology 14:155–179

Wu J, Spohn C (2010) Interdistrict disparity in sentencing in three U.S. District Courts. Crime & Delinquency

Yang CS (2014) Have interjudge sentencing disparities increased in an advisory guidelines regime-evidence from Booker. NYU Law Review 89:1268–1342

Yang CS (2015) Free at last? judicial discretion and racial disparities in federal sentencing. The journal of Legal Studies 44(1):75–111

Zatz MS (1987) The changing forms of racial/ethnic biases in sentencing. Journal of Research in Crime and Delinquency 24:69–92

Zhang J, Kai F (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. The Journal of the American Medical Association 280(19):1690–1691